COMMUNITY HEALTH SURVEYS
A Practical Guide for Health Workers

6. Presenting Survey Information

05099

Community Health Cell
Library and Documentation Unit
367, "Srinivasa Nilaya"
Jakkasandra 1st Main,
1st Block, Koramangala,
BANGALORE-560 034.
Phone : 5531518

Requests for copies of this publication can be made to one of the Regional Offices of the World Health Organization listed below.

Regional Office for Africa, P.O. Box No. 6, **Brazzaville,** Congo.

Regional Office for the Americas/Pan American Sanitary Bureau, 525, 23rd Street, N.W., **Washington,** D.C. 20037, USA.

Regional Office for the Eastern Mediterranean, P.O. Box 1517, **Alexandria** - 21511, Egypt.

Regional Office for Europe
8, Scherfigsvej, 2100 **CopenhagenØ** - Denmark.

Regional Office for South-East Asia, World Health House
Indraprastha Estate, Mahatma Gandhi Road, **New Delhi**-110002, India.

Regional Office for the Western Pacific, P.O. Box 2932, **Manila** - 2801, Phillippines.

COMMUNITY HEALTH SURVEYS
A Practical Guide for Health Workers

Other numbers in this series :

Number 1. Planning and Organizing
Number 2. Survey Sampling
Number 3. Using Available Information
Number 4. Questionnaire Design
Number 5. Interviewing and Recording

© International Epidemiological Association, 1989
Printed in Switzerland.

PRESENTING SURVEY INFORMATION

A Guide for Health Workers

Prepared for the International Epidemiological Association
in collaboration with the World Health Organization

by

Joycelin Chalmers
and
W. Lutz

Medical Statistics Unit
University of Edinburgh Medical School
Edinburgh, Scotland.

Published by

INTERNATIONAL EPIDEMIOLOGICAL ASSOCIATION

1989

TM-110
06098 N89
COMMUNITY HEALTH CELL
LIBRARY
AND
DOCUMENTATION

PRESENTING SURVEY INFORMATION

For whom :

1. Doctors and health services personnel responsible for providing specific services at local level and who need more information to improve or develop health services in their local community.
2. Doctors and health services personnel responsible for planning, administering or providing services in larger administrative units and who in the course of their work require information that is not already available.

Aims :

1. To enumerate and explain what needs to be done after the field work is completed and all the questionnaires have been checked. To achieve these aims, the discussion subdivides into four sections :
   (a) coding and information extraction
   (b) tabulation
   (c) basic statistical analysis and graphical methods
   (d) presenting and reporting the information
2. To emphasise the importance of an attractively presented and well written report. A survey remains unjustified unless followed by a widely disseminated report containing the main conclusions and which is later followed by appropriate actions to implement the survey recommendations.

Introductory Remarks

The methods of survey analysis and presentation discussed in this booklet are those most commonly used and which are easy to apply. The statistical procedures and graphical methods described will enable users to extract, by simple means, most of the information contained in the survey questionnaires and will enable users to present the information in a manner that is clear and easily understood.

In order to make the methods and computational procedures easy to follow, each is set out in systematic steps; full details of the computational and graphical procedures are given. The serious reader is encouraged to work through the examples for himself. Some computations and graphical methods are, unavoidably, slightly more complex and therefore have been given in the appendices so as not to distract readers by excessive detail on first reading.

Situations may arise where the analysis of the survey results requires additional or more advanced statistical methods. Some sampling designs may also necessitate more complex treatment. In such instances, the reader is encouraged to seek advice from statisticians or survey specialists. However, even in such situations, the reader who has understood the simple methods given in this booklet will find it easier to discuss his survey requirements with the specialists. Moreover, the advice given by the statistician will be far more meaningful to someone who has mastered the basic methods.

The previous booklets in this series, numbers 1 to 5, follow a special format where general principles are discussed on the left hand pages, whilst specific applications of the principles are illustrated and expanded on the right hand pages. In those five booklets the reader's attention is attracted by arrows to points of special importance between the general principles and their applications.

That format has **not** been adhered to in this sixth, and last, booklet in the series for two reasons :

1. Any general discussion of the principles of survey analysis and presentation must include fully worked examples so as to avoid becoming too abstract.
2. The discussion of the application of the methods of data handling, statistical computations and graphical presentation will repeat, of necessity, the methods already shown under the general discussion above. Repetition of basic methods of analysis and presentation, even if applied to an entirely different set of survey data, is likely to prove tedious and not very informative.

The format of booklet 6 is therefore that followed in most textbooks, the discussion flowing from one page to the next without distinct treatment of left and right hand pages.

Basic Procedures for Analysing and Presenting Survey Information

List of Contents

ANALYSING AND PRESENTING SURVEY INFORMATION

Section A : Coding

1. Introduction

Surveys are an important means, often the only practical way, of obtaining information about community and health conditions quickly and cheaply. The size of the study depends upon many factors. Some surveys consist of no more than 25 to 50 interviews. Small studies have the advantage of speed and economy; they present many fewer problems during the process of analysis. Larger studies, large in the sense of both the number of interviews as well as the number of facts, measurements and opinions recorded, provide more information, better coverage of the community and probably justify greater confidence in the stability and representativeness of the final results.

The size of the study affects the way in which the extraction of information, tabulation and analysis is organised. Very small studies, for example, do not gain much from coding the survey questions, whereas, for larger surveys, coding rapidly becomes indispensable.

Coding is given considerable emphasis for another reason also. Survey organisers, from the earliest days of survey investigations, have made much use of various punched cards, i.e. cards with holes punched in the card or along their edges, which then made possible the sorting and tabulation of the data by mechanical devices. These methods are no longer widely used, as they are being replaced by the methods provided by small microcomputers. Although useful small computers are not yet everywhere available, their cheapness is such that they are rapidly spreading to all parts of the world.

They are now available at many universities, high schools, technical colleges and government offices. Many who wish to undertake survey work are already, or soon will be, able to gain access to this new technology. If computers are to be used, then there is no alternative to coding all the survey information.

2. Coding Methods

Most survey questions are of the closed type* because of its general usefulness and convenience. As a rule, closed questions have been precoded on the questionnaire and further coding is then unnecessary. Occasionally this coding is not shown on the questionnaire and it must then be dealt with after the completion of the field work.

(i) Coding closed questions

Coding consists of choosing a symbol, most usually a number, to represent the answer given by the respondent. In the closed question the respondent is given a choice of options from which to choose. At coding, each of these options is then given a number, starting at 1, 2 and so on. A typical example from a completed questionnaire may look like this :

Question : How many people live in this house ?

One person	1. ☐
Two persons	2. ☐
3 to 5 persons	3. ☑
6 to 8 persons	4. ☐
More than 8	5. ☐
Don't know	6. ☐

* See Survey Booklet 4 : Questionnaire Design.

a) Closed questions whose options are mutually exclusive *

It is easy to devise a coding scheme for closed questions having mutually exclusive and exhaustive options.* Simply call the first possible response 1, the second possible response 2, and so on. In the above example the codes run from 1 to 6. In some situations it is necessary to distinguish between a "Don't know" answer, and a failure to get any response at all, i.e. when the question is left blank. Where this is necessary, a "Not answered/blank" code may be incorporated, so that the final code is :

Response Option	Code	Alternative Code
One person	1	1
Two persons	2	2
3 to 5 persons	3	3
6 to 8 persons	4	4
More than 8 persons	5	5
Don't know	6	8
Not answered/blank	7	9

The "Not answered/blank" category may appear in serveral questions, as will "Don't know". It is convenient, and certainly less confusing, if "Don't know" and "Not answered/blank" always have the same code for all questions in which they occur. One way to achieve this, is to give the last two categories the code 8 and 9, i.e. the codes 6 and 7 would not be used in this particular question, as is shown in the "Alternative Code" column above.

---

* See Survey Booklet 4 : Questionnaire Design.

In some studies it is unnecessary to distinguish between "Don't know" and "Not answered/blank" so that the two categories can be combined into a single "Don't know/Not answered" group and the single category can be given the code 9. For certain questions, it may also be necessary to allow for a "Not Applicable" code.

b) Closed questions whose options are not mutually exclusive

The most important point to make about a coding scheme is that it must be exhaustive, thereby making certain that every one of the survey questionnaires will have a code for the question concerned, including the questionnaires in which the question is left unanswered or is not applicable.

However, not all closed questions offer mutually exclusive response options. A question in which the respondent may choose more than one response in her * reply is often referred to as a multiple response question.

The above can be illustrated by an example from a survey dealing with conditions and work loads of junior doctors in teaching hospitals. The responses of two junior doctors are given below.

---

* In keeping with the previous survey booklets, the respondent is assumed to be a woman, although in practice, the respondent might just as easily be a man.

Question : "During the past seven days, which of the following duties did you perform ?"

Response Options :	**Response Code***	**First doctor**	**Second doctor**
Caring for patients whose main problem was gastro-intestinal	1	☑	☐
Caring for patients whose main problem was **not** gastro-intestinal	2	☑	☐
Working in the casualty/ emergency department	3	☐	☑
Teaching students or doing research	4	☐	☐
Hospital activities not directly connected with any of the above	5	☑	☑
Not on duty for any of the past seven days or otherwise not applicable	9	☐	☐

Such questions can be coded in several ways, two of which we will consider. The coding method chosen should :

1. facilitate (help) the kind of statistical analysis planned for the question.
2. be as simple as possible, subject to meeting the first condition.

---

* The Response Code is also referred to as the Question Code or simply as the Code.

First Approach

Occasionally, interest centres on how often a particular option was ticked, irrespective of how many others were also chosen. If the statistical analysis is restricted to this simple approach, then it is sufficient to construct a coding system exactly like the mutally exclusive type of closed question; this coding scheme is shown in the above example and consists of the codes 1, 2, 3, 4, 5 and 9. It is the simplest possible coding scheme for multiple-response questions and is well suited if we are only interested in simple information such as : "How many of the doctors interviewed were on casualty or emergency duties during the past seven days ?"

To answer the question, it is necessary only to count how many doctors ticked the third option, irrespective of whether they also ticked other options. However, if we wish to know how many of the doctors interviewed had done casualty duties as well as cared for other patients, then it is necessary to select those questionnaires on which options 1 and/or 2 as well as 3 have been ticked.

Although it can be done, the selection is tedious and liable to clerical errors.

Second Approach

For questions whose response options are not mutually exclusive and for which combinations of responses, i.e. patterns of response, are to be studied, the following two coding schemes should be considered.

Coding Method 1

The following method of coding multiple-responses is strongly recommended where there are only two or three responses to the question.
For example :

What kinds of milk are provided for your baby ?

(a) Breast milk ☐
(b) Cow's milk ☐
(c) Other, including tinned and powdered ☐

If these are the only three options offered, the possible pattern of responses are :

Pattern	Suggested Code
(a) only	1
(b) only	2
(c) only	3
(a) and (b)	4
(a) and (c)	5
(b) and (c)	6
(a), (b) and (c)	7

If "Not answered" or "Not applicable" are appropriate, then codes 8 and 9 can be added.

When there are four or more response options, the number of possible response patterns becomes large and requires at least double figure numbers. This can be demonstrated by extending the options from 3 to 4 in the above question.

(a) Breast milk ☐
(b) Cow's milk ☐
(c) Goat/sheep milk ☐
(d) Tinned or powdered milk ☐

The possible patterns are now :

	Code		Code		Code
(a) only	01	(a) + (d)	07	(a) + (c) + (d)	13
(b) only	02	(b) + (c)	08	(b) + (c) + (d)	14
(c) only	03	(b) + (d)	09	(a) + (b) + (c) + (d)	15
(d) only	04	(c) + (d)	10	Not Known/Not answered	88
(a) + (b)	05	(a) + (b) + (c)	11	Not applicable	99
(a) + (c)	06	(a) + (b) + (d)	12		

Note how, in the above case, where the code goes into double figures, the code for "Not known/Not answered" becomes 88, instead of 8. Likewise the "Not applicable" code is written as 99. Note also that all the codes are expressed in double figures, e.g. the first code is written as 01 and not just as a simple 1.

The code can be simplified if interest centres on only a few of these patterns. In such a case, all the remaining patterns of lesser importance to the study can be grouped into one single category "Other", thus reducing the number of codes required.

However, once there are as many as five or more response options, Method 1 becomes involved and difficult; it may then be better to consider the alternative approach, discussed under Method 2.

Coding Method 2

Step I : On a sheet of ruled paper draw as many vertical columns as there are response options in the question. The first column is set aside for the first response option, the second column for the second response option, and so on.

Step II : For each response option, enter a 1 in its corresponding column if it was ticked; enter 0 if it was not chosen. This procedure constructs a six digit code corresponding to the respondent's choice of options.

A typical coding sheet for the above survey question then looks like the following :

	Question Code						
Respondent	**1**	**2**	**3**	**4**	**5**	**9**	**Derived Code**
First Doctor	1	1	0	0	1	0	110010
Second Doctor	0	0	1	0	1	0	001010
Third Respondent	0	0	0	0	0	1	000001
Fourth Respondent	0	1	0	1	1	0	010110
Fifth Respondent	1	1	0	0	1	0	110010
Sixth Respondent	0	0	1	1	1	0	001110
Seventh Respondent	1	1	0	0	0	0	110000
Column Totals	3	4	2	2	5	1	

By totalling each column separately, it is immediately seen how many of the doctors surveyed engaged in each of the possible duties. The scheme also allows enumeration of doctors who engaged in particular combinations of duties.

Of course, even this type of coding is tedious and prone to error if done clerically. Double checking, by another person, is very necessary. Such coding becomes particularly advantageous where workers have access to a computer.

c) Coding of priorities and pathways

In the survey dealing with working conditions of junior doctors the question to junior doctors : "During the past seven days which of the following duties did you perform ?" was immediately followed by the question "Please tell me which of these duties took up most of your time, which second most, and so on ?". The interviewer was asked to record the replies by entering the rank order of each response chosen when answering the first of these two questions. The answers for the first and second doctors were as follows :

	Code	First doctor	Second doctor
Caring for patients whose main problem was gastro-intestinal	1	2✓	☐
Caring for patients whose main problem was **not** gastro-intestinal	2	1✓	☐
Working in the casualty/emergency department	3	☐	1✓
Teaching students or doing research	4	☐	☐
Hospital activities not directly connected with any of the above	5	3✓	2✓
Not on duty for any of the past seven days or otherwise not applicable	9	☐	☐

These two questions are a typical example of respondents being asked to choose from a list of options and also to order (rank) their choices as to which they would do first, or which was of greatest importance to them, which came second and so on. Respondents are in effect asked to indicate the path or their priorities in their responses to the question.

The coding scheme for the above questions is required to reveal **both** the options chosen and the order in which the respondent places her options. A suitable coding scheme can be devised as follows :

Step I : As in the previous section, on ruled paper draw vertical columns, one column for each of the response options given in the initial question. The first column corresponds to the first option, the second column to the second option, and so on.

Step II : In each column, enter the corresponding rank given by the respondent. Options not chosen are given the value 0 (zero). The required code is the sequence of numbers thus created, from left to right.

As an illustration, the codes for the first and second doctors are given by :

Respondent	Question Code						Derived Code
	1	2	3	4	5	9	
First doctor	2	1	0	0	3	0	210030
Second doctor	0	0	1	0	2	0	001020
Third doctor etc.							

A word of warning. The codes are complex and require careful analysis; they should not be used unless really necessary. However, they can be most informative and useful, particularly for large studies where pathway analysis without the help of such coding becomes difficult. Simple analysis of such codes is done by counting the number of 1's, 2's, 3's and so forth, in the first column of the code, thereby showing how many respondents gave first priority to the activity (as coded in the first column), how many gave the activity second priority, and so on. By counting also the number of zeros we see how many respondents did not engage in the particular activity. Repeating the process for each column of the code in turn, a similar analysis is obtained for each of the coded activities.

Analysis of this type of coding becomes easier if a computer and suitable computer programs are available.

Whilst the actual construction of closed question codes is quite straightforward, the above discussion makes clear why survey planners are advised, whenever possible, to formulate closed questions with mutually exclusive response options. When multiple responses to a question are unavoidable, the number of combinations of options soon becomes large and the resulting codes unwieldy and hence more troublesome to check and analyse.

(ii) Coding open questions

The open question is answered by each respondent in her own words and what she says is recorded on the questionnaire. Each respondent's reply will be different in some respects from all the other replies; there are no pre-set response options for an open question.

Basically, only two things can be done with open questions and these are :

1. summarise or quote a few of the respondents' answers as examples in the report, to illustrate some aspect of the study, i.e. use only a few replies as examples of what the respondents said and felt.
2. devise a coding scheme that will allow information from open questions to be extracted, tabulated and analysed in the same way as for closed questions.

At the coding stage it is often realised, despite earlier admonitions (warnings) to exclude unnecessary questions, that some of the questions are unlikely to be analysed because they are not central to the objectives of the study. If possible, such inessential questions should now be identified, particularly if they are open questions, because of the time and effort required in their coding.

Procedure for coding open questions

Although all respondents express themselves differently in an open question, the meaning of their answers, the reasons they give, the objections they raise, or the agreements they express will often be rather similar. Coding open questions makes use of this similarity.

The most taxing (difficult) aspect of coding open questions is the construction of generic* categories that summarise general aspects of the respondents' answers. The generic responses will usually not be mutually exclusive;

* A generic category is a category such that similar or related answers can all be included within the same class.

however, the coding scheme discussed below will produce exhaustive generic categories, i.e. each respondent's answer will fit at least one of the codes.

As an illustration, in a study of why, when and how parents bring their ailing (sick) children to the Health Centre, one question, for appropriate cases, was :

> "What were your reasons for not bringing your child to the Health centre sooner ?"

Typical summary responses to this question, called generic responses, included :

1. Did not at first realise the seriousness of the illness.
2. First sought other medical aid, including traditional medicine.
3. No one available to bring the child, including ill-health of parents.
4. Financial or transport difficulties, including lack of money, poor transport, long distances, etc.
5. Other reasons.
9. Not answered/not applicable.

Many of the respondents' answers included one or more of these generic responses, although each respondent had expressed her reasons in different words.

The following simple method for coding open questions has been used by epidemiologists at W.H.O.

Step I : Take a random sub-sample of questionnaires. Twenty-five or so will usually be sufficient, although for large studies a bigger sub-sample is advised.

Step II : Study the answers to the question in this sub-sample and, from these, construct short summary responses, i.e. generic responses, that are typical of the answers given.
A respondent's answer may of course contain within it more than one generic response. This is to be expected. because, in their reply to a question, respondents often give a combination of reasons or give serveral distinct bits of information, each of which may belong to a different generic category.

Step III : Count the number of respondents (in this sub-sample) whose answers fit each of the generic categories. A simple tabulation is often the easiest way of doing the counting.

Step IV : Rank the generic categories by the number of answers in each. Code from 1 to 8 the top eight ranking groups. Code as 9 those questionnaires that have no answer for the question. Finally, give code 0 for all other answers that do not fit any of the geneirc categories, coded 1 to 8. (If there are less than 8 generic categories, the "Other" category should preferably not be coded as zero).

Step V : As a check, select a further ten or so questionnaires at random and see whether this coding scheme also works well for these questionnaires. If the coding scheme does not work well, then revise the wording and content of the generic categories and repeat the process.

In coding open questions, two further points should be kept in mind :

1. The number of generic categories should be kept to a minimum. Excessively fine generic responses rarely add much useful information, but will inevitably add to the complexity of later tabulation and analysis. Thus under Step IV, it is preferable to have less than eight generic categories if the questions and the responses given permit this.

2. Having set up the coding scheme as outlined, it sometimes happens that the "Other" category becomes very large, which may suggest that the "Other" group contains within it generic responses that are worth isolating and treating as separate categories. It is worth checking by having a closer look at the type of answers grouped together under this "Other" heading. It may of course happen that one of the other generic groups becomes very large, thereby suggesting that this response category might also beneficially be divided into two or three separate response groups. Whether or not it turns out to be the case will depend upon the circumstances of the study. However, the aim should be to keep the number of response categories small and not to subdivide response categories unless there are good reasons for so doing.

Some open questions are so general that the number of generic response options is unavoidably large, more than 9 or 10. In such questions, the coding scheme must allow for the large number of responses and it then becomes necessary to have two figure codes, starting at 01, 02 and ending with 99 for "Not known". Such a detailed coding scheme can only be justified for large studies as otherwise many of the categories will contain only very few entries.

After completing the coding for all the questions, the questionnaires are ready for sorting into suitable piles, in preparation for the next stage in the analysis of the survey results.

3. Sorting the Questionnaires

Preliminary to extracting and analysing the survey information, it is necessary to sort the questionnaires into convenient piles or groups. The way in which the sorting is done depends primarily on the kind of sampling scheme used for the study, the reason being that the graphs and statistical calculations may be done differently, depending on the sampling method. Hence the sorting of the questionnaires must be so arranged as to assist doing these calculations later on.

In the booklet on Survey Sampling*, five different sampling schemes were discussed, and a brief description of these is given in Appendix 1. To understand the discussion that now follows, the reader should refresh his or her memory of what the sampling schemes are.

Fortunately, for purposes of sorting, these five sampling schemes can be grouped into just three types :

---

* Booklet No. 2 in this series.

(1) List Sampling (2) Numbered Tag Sampling	Sorting Method A
(3) Stratified Sampling	Sorting Method B
(4) Cluster Sampling (5) Two Stage Sampling	Sorting Method C

Each of the above three sorting methods is described below.

There are, of course, many other survey sampling schemes besides the five listed here. If a survey design has been used that is not one of the above five, then the method of sorting the questionnaires will almost certainly need to be changed accordingly; advice shoud be sought either from the person responsible for the sampling design or from a statistician.

The Sorting Methods B and C, about to be described, depend on each questionnaire having the strata*, clusters* or other appropriate identification information clearly recorded. If this information has not been recorded, it will not be possible to analyse the survey returns according to proper statistical principles.

---

* See Appendix 1 for a definition of these terms.

Sorting Method A : Suitable for List or Numbered Tag Sampling Designs

No systematic sorting is required. The questionnaires can be neatly arranged in conveniently sized piles and the analysis can proceed to the next stage, which is the extraction of information, a process often referred to as data extraction.

Sorting Method B : Suitable for Stratified Sampling Designs

The questionnaires must be carefully separated into groups, according to which stratum they belong. There must be a separate pile of questionnaires for each of the survey strata.

Sorting Method C : Suitable for Cluster and Two Stage Designs

Separate the questionnaires into groups, according to which cluster they belong. There must be a separate pile of questionnaires for each of the clusters drawn into the survey sample.

Checking : the sorting of questionnaires, whether for method B or C, must be checked, preferably by a second person. It is most important that each questionnaire is placed into its correct pile.

4. Extracting the Information

Data (information) extraction from the survey questionnaires is basically a manual * (clerical) job, but it must be done with intelligence, concentration and care.

Two methods of extraction are in common use :

(i) tally chart extraction

(ii) summary sheet extraction.

(i) Tally chart extraction

The tally chart is a simple and direct method for finding the distribution of values for any given characteristic. As an example, a survey to determine the number of children under 5 years who have been immunised might record values such as : 1, 0, 2, 0, 0, 1, 3, 0, 0, 0, 2, ..., each of these values appearing on a separate questionnaire in response to the question :

> "How many of the children under 5 years, in this house, have been immunised ?"

To get a better idea of what these data look like, they should be tabulated; the steps necessary are as follows :

Step I : Decide on suitable intervals or values for the table.

Step II : For each questionnaire, place a tally stroke (short line) next to the corresponding value, or interval, in the table. Every fifth stroke is put horizontally to make counting easier, e.g. six strokes is better written down as : 𝍸 I

---

* If a microcomputer and suitable computer programmes are available then the processing of the data is done differently from this stage onwards. Nevertheless, the aims and requirements of the analysis must still be formulated by the survey organser. Hence, even with microcomputers, Part 4 of Section A and the following Sections will still be relevant.

Step III : Count the number of strokes against each value or interval of the table; record also the grand total, i.e. the sum of all the separate counts, for the whole table.

Step IV : Repeat the process a second time, comparing the second result with the first to ensure there are no errors.

With the above illustrative data, the procedure would start as :

No. of children under 5 years immunised per household	Number observed (tallies)
0	𝍸 I
1	II
2	II
3	I
4	
5 or more	
Not known/ Not answered	
Total :	

The final table, after all the questionnaires have been entered, might then look like this :

No. of children under 5 years immunised (a)	Tallies of Households (b)	Number of households (HH) (c)	Total number of children immunised (a) x (c)
0	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 III	33	0
1	𝍸 𝍸 𝍸 I	16	16
2	𝍸 II	7	14
3	IIII	4	12
4	I	1	4
5 or more		0	0
Grand Total		61	46
Not known/ Not answered	II	2	
Total HH visited		63	

There are thus 61 households responding out of the 63 visited.

The total number of children under 5 years immunised in the 61 households for which we have a response, is then found by multiplying, row by row, the number in column (a) by the number in column (c) and totalling (adding) the results. The calculations are shown in the last column of the above table, i.e. 46 children under the age of 5 in these 61 households have been immunised.*

Note : All tables should show the number of cases for which an answer is not available. In the above example, there are two households for which the immunisation data could not be obtained.

* To derive indices of immunisation, it is also necessary to have information on the number of children under 5 years in these same households. This is illustrated and discussed on pages 35 and 36.

The procedure is similar if, instead of single specific values, the table is to display class intervals. For instance, if a Health Centre wishes to study the live birth weights of children delivered at the Health Centre or by its midwives during the last 12 months, then a suitable table might be :

(a) Birth weight in kg.	(b) Tallies	(c) Frequency	(d) Mid Point Value of (a)	(e) Multiply (c) x (d)
under 1 kg		0	0.50	0.0
1.0 < 2.0*	I	1	1.50	1.50
2.0 < 2.5	𝍸	5	2.25	11.25
2.5 < 3.0	𝍸 𝍸 III	13	2.75	35.75
3.0 < 3.5	𝍸 III	8	3.25	26.00
3.5 < 4.0	III	3	3.75	11.25
4.0 < 4.5	I	1	4.25	4.25
4.5 < 5.0		0	4.75	0.0
Grand Total		31		90.00

* Symbols such as 1.0 < 2.0 are read as one kilogram but less than two kilograms and similarly for the other intervals. Thus a baby of exactly two kilograms is entered as belonging to the third interval, but a child of exactly 2.5 kg. would be entered into the next, the fourth interval.

To calculate the total live birth weight of all the babies, the mid point value of the intervals is used. It is calculated in column (d). The total weight of the 31 babies is then the sum of (c) x (d), as given in the last column, (e).

The average live weight for the 31 babies is found by dividing the total weight in column (e) by the number of babies = 90/31 = 2.94 kg. *

* The average, and other statistical indices used here for illustration, will be discussed in a later section.

The tally chart procedure is restricted in its usefulness to the extraction and tabulation of single characteristics, one at a time. It is clearly inadequate for all but the simplest kinds of analysis.

If, in the immunisation example given earlier, it was required to calculate the proportion or percentage of children immunised, it is necessary, using the tally chart approach, to go through all the questionnaires twice, first to find the total number of children under five, immunised or not, and then to find the number who were immunised. Going through questionnaires, time and time again, is a slow business and should be avoided where possible.

The usual way around this difficulty is to construct a summary chart as an intermediate step between extraction and tabulation.

Note : The tally stroke extraction discussed above applies only to surveys for which sorting Method A is applicable, i.e. for surveys using the "List" or the "Numbered Tag" sampling schemes. For surveys for which sorting Methods B or C apply, a separate tally stroke extraction should be done for each of the strata or clusters. This point is again stressed on page 37.

(ii) Summary chart extraction

The summary chart consists of a ruled sheet on which the name of the study unit *, usually the name of a patient or identification of a household, is written on the page. Across the page are the variables, i.e. the measurement and the characteristics, that have been recorded. A simple summary chart for the above immunisation example would appear as follows :

---

* The study unit is the basic or smallest unit with which the survey is concerned; it is this unit that the field workers must ultimately visit for interviewing, inspection or study.

Characteristics and Values

Study Unit (HH)	Male * persons in HH	Female* persons in HH	Children under 5 yr. in HH	Children under 5 yr. immun.	No. of children aged 5 to 16	No. of etc. children aged 5 to 16 at school
(1) Desai, J Gandhi St.	3	4	2	0	2	2
(2) Tooli, H Bridge St.	4	3	1	1	2	1
(3) Dadoo, J Panda St.	2	5	3	1	2	2
(4) Krishna, F India Terr.	3	1	1	1	1	1
(5) Singh, K Bangor Dr.	3	4	2	2	2	2
(6) Kanji, G Delhi St.	1	1	3	0	1	1
continuation (7) –	–	–	–	–	–	–
–	–	–	–	–	–	–
Totals :	117	129	59	46	62	51

* All ages

Constructing the summary chart is very simple. The values and characteristics of each study unit are written out across the page. When the first study unit's summary has been completed, that questionnaire is set aside and the next unit's questionnaire is entered onto the summary sheet, in a similar manner, until all the questionnaires have been dealt with.

In theory, it is only necessary to go through each questionnaire once to extract all the information onto the summary sheet, thereby speeding up the data extraction immensely.

In practice, there are often too many variables (measurements and characteristics), to be able to do the analysis in one perusal (examination) of the questionnaires. As a result, several summary sheets are usually drawn up, each continuing from where the previous summary sheet left off; **the study unit's identification number must, however, appear on every summary sheet,** making the process take a little longer. Even so, the summary sheet offers a considerable saving in time and effort.

The column grand totals at the bottom of each summary sheet are very useful and will often, without further effort, allow the calculation of important statistical indices. For instance, from the above :

(1) the ratio of females to males (i.e. the sex ratio)

$$\frac{129}{117} = 1.10$$

(2) the percentage of females in the sample

$$\frac{129}{(117 + 129)} \times 100 = 52.4\%$$

(3) the percentage of children under 5 years who have been immunised :

$$\frac{46}{59} \times 100 = 78.0\%$$

(4) the percentage of children of school-going age (5 to 16 years) who are attending school

$$\frac{51}{62} \times 100 = 82.3\%$$

A most important point to remember is that the summary chart must correspond to the sampling scheme used in the study.

As previously shown, the questionnaires are first sorted into piles (groups) appropriate to the sampling design. A separate summary sheet must be drawn up for each of the piles and used only for its own particular pile of questionnaires. The summary sheet must therefore have clearly written on it the pile to which it refers. The reason for the careful separating out, i.e. a separate summary sheet for each separate group of questionnaires, is that the summary sheet totals cannot simply be added together for all sampling designs. In some sampling schemes, the summary sheets need to be combined arithmetically according to procedures that are appropriate to the sampling design.

5. Checking for Errors

Before proceeding to the next stages of tabulation, calculating statistical indices and drawing graphs, it is essential that the copying of the data from the questionnaires onto the summary charts has been done without mistakes. Copying, especially if there is much of it, is a common source of errors in statistical work. If the summary charts have errors in them, then the mistakes will appear in, and affect the tables, calculations and graphs derived from the summary sheets.

The only effective way of checking the summary chart is to have a second person draw up similar summary sheets. The two sets of summary sheets must then be compared. If any discrepancy (difference) is found, further checking is required until the error is located and put right.

Section B : Tabulation

1. Planning the Statistical Analysis

There are two separate aspects to the planning of the statistical analysis of a survey. They are :

(a) Deciding upon the tables required and the type of statistical methods to be applied in order to answer the questions posed (asked) by the study.

(b) Organizing the whole process of coding, data extraction and tabulation and applying to this data the appropriate statistical methods, including also the necessary checking. Decisions are required as to who is to do this work, in what order, and the time the various tasks require, allowance being made for necessary supervision and the instructions/guidance needed by the assistants.

(i) Choice of statistical methods

Normally, it is only possible to decide upon the broad requirements. As the statistical analysis proceeds new ideas will emerge that suggest additional methods should be applied and that some of the data should be explored in greater depth. Readers should also be warned that new ideas may suggest finer or different breakdowns of the data, requiring new summary tables. This is a rather common situation at the outset of analyses, and can be disastrous for hand tabulation. Nevertheless, it is important, for an efficient analysis, to be clear from the outset (start) as to the main methods required and to which survey questions the methods are to be applied.

The majority of the statistical analyses are required to provide the tools (methods) to :

(a) describe the situation as found during the survey.
(b) compare the study results with other surveys or with data available from other sources, such as regional and national figures.
(c) study the relationship existing between some of the variables recorded during the survey.
(d) study trends and changes over time.

There are several simple statistical methods that can be fruitfully used for all four of the above categories. They include calculating :

1. Averages
2. Medians
3. Proportions
4. Percentages
5. Ratios

In addition, there are the methods of contingency tables whose uses are generally confined to categories (b), (c) and (d) above and less so to (a), and correlations which are largely restricted to categories (c) and (d).

Certain simple graphical methods also have wide and useful applications. The pie chart and histogram are most often used when dealing with categories (a) and (b) type situations. The scatter diagrams are most frequently employed when studying relationships between variables, whilst time charts display trends and changes over time.* However, these methods are used for all four categories, under appropriate conditions. The methods and conditions under which they may be used will be expanded upon later.

---

* See pages 75-78 for examples.

(ii) Organising the process of analysis

The analysis of the survey questionnaires is usually a protracted (long) process that needs to be planned and organised before the fieldwork has been completed. The planning of the analysis should consider at least four aspects :

1. Estimate the resources and the time needed to carry out the analysis. This is not always easy, and even survey specialists sometimes produce incorrect estimates. Nevertheless, make some estimate and, if in doubt, allow for more resources and time, rather than less.
2. Draw up a schedule of the various stages of the analysis and the times (dates) by which these should be completed.
3. Consider the skills and the knowledge needed by the persons doing the clerical sorting, coding, data extraction and other tasks, in order to do the work properly. Plan instruction and training sessions, where these are considered necessary, well before the work is to start. Give some thought also to the supervision and monitoring of the analysis whilst it is in progress.
4. Decide on priorities. Some aspects of a survey may need to be analysed more quickly than others because the results and conclusions of those sections must be known as soon as possible. Arrangements must then be made to give priority to particular sections and to ensure the priorities are maintained.

2. Tabulation

As already emphasised, separate summary charts should be drawn up for each of the groups into which the questionnaires have been sorted. In surveys in which a List or Numbered Tag sampling scheme was used, there is only one group of questionniares, the entire lot. In other types of sam-

pling, each group and its corresponding summary sheets must be kept separate. Provided the number of questionnaires in a pile (group) is large enough, say 20 or more, it is then worth while constructing separate tables for each group.

Note :

The discussion to follow applies only to surveys in which either List or Numbered Tag sampling procedures have been used. Appendix 2 will describe how summary sheets and tables for other sampling designs, i.e. not List or Numbered Tag sampling, can be combined to provide an overall picture and estimates for the community as a whole.

(i) Frequency tables

The most commonly used of the statistical tables are frequency tables. They are often referred to as distribution tables as they display how the sample values are distributed, thereby allowing important estimates to be made about the community from which the sample was drawn. Moreover, frequency tables are particularly helpful in comparing survey results with similar data obtained from elsewhere.

How to derive the frequency table can best be described by using actual examples. The data to be used here for illustration are taken from several different surveys and regional census figures. In particular, use is made of a selection of data collected on heart disease in Scotland during a community survey * that involved the study of 448 men aged 45 to 54. To simplify the illustration only the results of the first thirty cases are given below :

---

* Edinburgh-Fife Heart Study (1980). Principal investigators : Professor M.F. Oliver and Dr. Mary P. Fulton.

Summary Chart

Respondent Number	Age (years)	Height (cm)	Weight (kg)	Syst. B.P.* (mm Hg)	Diast.B.P.** (mm Hg)
1	50	173	72	125	78
2	50	180	79	**105**	72
3	47	169	68	133	82
4	52	158	69	153	84
5	51	179	93	117	70
6	48	169	70	142	85
7	49	162	67	**195**	116
8	52	178	77	134	87
9	49	167	66	119	69
10	50	168	74	120	75
11	55	166	105	144	81
12	46	168	74	128	87
13	46	182	103	135	76
14	53	170	71	118	78
15	51	179	75	137	82
16	52	167	78	151	85
17	48	178	92	148	84
18	52	173	92	121	74
19	50	166	70	134	77
20	53	171	79	151	85
21	55	166	78	146	76
22	54	174	79	141	94
23	55	172	83	169	101
24	50	163	81	152	88
25	54	172	61	138	87
26	46	174	62	140	89
27	54	173	84	156	98
28	49	166	81	152	99
29	51	181	78	124	77
30	48	183	83	114	81

* Systolic blood pressure
** Diastolic blood pressure

(ii) Single variable frequency tables

The single variable (one-dimensional) frequency table describes the distribution of a single characteristic or variable and can be constructed as follows :

Step I : Scan through the values in the summary chart to find the minimum and maximum values in the sample, i.e. find the range.

Step II : Divide the range into a convenient number of intervals.
Between four and twelve intervals is usually the most practical number.

Step III : Use the intervals as the class intervals for the frequency table and by the tally stroke method determine how many of the sample values fall into each of the class intervals.

Example : The distribution of systolic blood pressures as recorded in the sub-sample of 30 cases from the Edinburgh-Fife Study.

Step I : The maximum and minimum systolic blood pressure in this sub-sample are 195 and 105 respectively, i.e. a range of 90.

Step II : As the sample size is small, consisting of only the first 30 cases, a table of five or six intervals seems advisable; a class interval size of 20 is therefore suitable and convenient. A useful practical hint (suggestion) is to start the first interval a few points below the minimum value. A starting point of 100 has been taken for this sample.

Step III : The frequency table, using the suggested intervals is then given by :

Sys. B.P. (mm Hg)	Tally Strokes	Frequency	Percentage * frequency
100 - 119	𝍸	5	17
120 - 139	𝍸 𝍸 |	11	37
140 - 159	𝍸 𝍸 ||	12	40
160 - 179	|	1	3
180 - 199	|	1	3
		30	100

(iii) Comparison of frequency tables

There is often a need to compare tables of similar data, but derived from different sources, such as comparing a table based on survey data with national or regional tables. Such comparisons are facilitated (made easier) if :

(a) each of the tables is based on a sufficiently large number of cases, preferably not less than 50.

(b) the class intervals of the tables are the same.

(c) the total number of cases (grand total) for each table is the same.

The reason for the first requirement, that the totals for each table should be reasonably large, preferably 50 or more, is that survey data is based on a **sample.** Experience makes us realise that sample results are variable, i.e. if a second but

* As a general rule, it is unwise to calculate percentages where the total is less than 30; some would say less than 50.

otherwise very similar survey were done, then the findings would not be identical to those obtained during the first study. There would be small variations and differences in the figures and values obtained each time the study was repeated. These variations between studies will become relatively less important the larger the survey, i.e. the larger the sample size, unless there have been substantial changes in the community during the time elapsed between the surveys. Hence, when comparing studies or statistics from other sources, it is important to be confident that the tables are based on sufficiently large numbers to be stable and reliable.

The last requirement, i.e. to have similar grand totals, is seldom satisfied. In order to simplify the comparison when the sample sizes are unequal, we can convert the tables into percentage frequency tables.

(iv) Percentage frequency tables

Only two steps are needed to convert a frequency table to a percentage frequency table.

Step I Divide each class interval frequency by the grand total for the whole table and multiply the result by 100. This is equivalent to (the same as) calculating the percentage of the total that falls into each of the class intervals.

Step II Check the artihmetic by ensuring that the addition of all the class interval percentages equals 100, or very close to 100, as a small rounding error may be unavoidable. Normally, a total between 99.9 and 100.1 is acceptable. As a rule it is undesirable to express percentage to more than one decimal place; a percentage of, say, 13.6 is acceptable but 13.589 suggests a very misleading precision.

It is better for many purposes, and particularly if the grand total is less than 100 or so, to round the percentages to the nearest integer (whole number). Thus a percentage of 13.589 would be rounded to the nearest whole number and simply recorded as 14 per cent.

Where percentages are rounded to the nearest integer, it may happen that the sum of the percentages is 99 or 101. Many statisticians adjust the largest of the percentages up or down by 1, so that the sum is exactly 100.

(v) Two-dimensional frequency tables (Contengency tables)

Two-dimensional frequency tables are often referred to as **contingency tables.** Contingency tables consist of a square or rectangle divided into rows and column boxes, called cells. The rows correspond to one variable * and the columns to some other variable that is thought to have a connection with, or have a bearing on, the row variable. Thus contingency tables are designed to study the relationship between two variables such as height and weight in children, diastolic and systolic blood pressures, or age and myopia.

The stepwise construction of a contingency table proceeds as follows :

Step I : Find the maximum and minimum for each variable and decide upon the appropriate number of class intervals wanted, usually between four and twelve intervals. The two variables need not have the same number of intervals.

---

* A variable is a generic term for any of the numerous measurements and characteristics such as age, height, blood pressure, number of children, and so on recorded during the study.

Step II : Using the summary chart, enter a tally stroke for each case (individual or sampling unit) into the box that corresponds to the two values of this case. Cases that do not have a value for each of the two variables, must either be omitted from the table or entered into a row or column set aside for unrecorded values.

Step III : Add up the number of tally strokes in each box, in each row and in each column. The sum of all the row totals **must equal** the sum of all the column totals because each of these should equal the grand total of cases in the contingency table. If the totals are not the same, one or more errors in addition have been made.

Step IV : Repeat the whole process and compare the two contingency tables. Step III only checks whether the additions have been done correctly whereas step IV checks that the extraction from the summary table is correct.

The above simple process can best be illustrated by an example using the previously given 30 cases from the Edinburgh-Fife Heart Study. The contingency table for the relationship between diastolic and systolic pressure is obtained as follows :

Step I : Range for diastolic B.P. : 69 to 116
Range for systolic B.P. : 105 to 195
In such a small sample, only 30 cases, five or six intervals for each variable seems about right.

Step II :

(i) Construct the rows and columns corresponding to the class intervals chosen in step I. Add one more row and one more column for the Not Recorded or Not Known cases.

(ii) Transfer the data from the summary chart to the contingency table, using tally strokes.

Step III : Check row and column totals.

Systolic B.P. (mm Hg)	Diastolic B.P. (mm Hg)							Row Totals
	65-74	75-84	85-94	95-104	105-114	115-124	Not Recorded	
100-119	III	II						5
120-139	I	~~IIII~~ III	II					11
140-159		IIII	~~IIII~~ I	II				12
160-179				I				1
180-199						I		1
Not Recorded								0
Column Totals	4	14	8	3	0	1	0	30

Note :

1. The row and column totals separately add up to the grand total of 30; this is a check that must **never** be omitted.
2. In this particular table there are no entries in the "not recorded" boxes (cells) because a systolic and diastolic pressure was recorded for each of the 30 cases.

(vi) Interpreting contingency tables : What to look for

The construction and interpretation of the contingency table is a good method for exploring survey data because :

(a) it automatically provides the frequency table for each of the two variables used, i.e. the row and column totals, often called "marginal totals", are the respective frequency tables. In the above example the row and column totals give the frequency tables of the systolic and diastolic pressures respectively.

(b) if there is a pronouced (strong) association or relationship between the two variables, then the association can be seen from the cell frequencies. Where both variables tend to have values in the same direction, i.e. where one of them is large the other tends to be large also, then the central boxes (diagonal boxes) will show the biggest cell frequencies. When the association is in the opposite direction, i.e. when one variable has a large value there is a tendency for the other variable to have a small value, then this too is shown by the concentration of frequencies in the diagonal boxes, but in the opposite direction than before. If there is no association, or only weak association, between the variables, then the cell frequencies are distributed more widely throughout most of the cells with little or no concentration along the diagonal cells.

In the above systolic-diastolic B.P. table, the diagonal cells have the larger frequencies, showing clearly that :

(i) a strong association exists between systolic and diastolic pressures

(ii) the association is in the **same** direction, i.e. with large systolic pressures there is a tendency for the diastolic pressure to be high also.

The contingency tables (for similar variables) can be compared visually, provided that :

(a) the sample size is sufficiently large, preferably 50 or more cases

(b) the class intervals used for the variables are the same for both contingency tables

(c) in each contingency table, the frequencies have all been expressed as percentages of the grand total.

Note :

More efficient statistical methods for comparing contingency tables exist than are described here, but the reader is referred to statistical texts for their use.

Section C : Statistical and Graphical Methods

1. Basic Statistical Estimates

The principal reason for calculating statistical estimates is to assist the research worker to generalise from the survey results to the whole community from which the sample was drawn.

The following are amongst the most useful and simple of the statistical estimates *

(a) representative sample values :
    (i) the average (mean)
    (ii) the median

(b) proportions and percentages

(c) sample ratios

(d) community totals

(e) variability :
    (i) the range
    (ii) the quartiles
    (iii) the standard deviation.

However, the above statistical estimates may be calculated by different artihmetical procedures depending on the survey sampling design. The method of calculation in this section is only applicable to simple random sampling, two examples of which are the List ** and Numbered Tag ** sampling schemes described earlier. The manner in which these calculations can be modified to meet the needs of the other types of sampling is described in Appendix 2.

---

* There are, of course, many other useful estimates that epidemiologists and statisticians can use to interpret and to generalise from the survey information, reference to which can be found in textbooks on statistics, especially those devoted to survey methods.

** See Appendix 1.

There are two estimates that are commonly used to provide a typical or representative value about which the individual sample values will fluctuate, some being larger and some smaller than the typical value. None of the sample values need coincide with the typical value, although the sample may contain such values.

The two estimates, also called "measures of location" because they locate the representative values, are :

(i) the average (also called the mean)

(ii) the median.

The two indices will not in general be the same, but each is in some sense representative or typical of the sample data. Although the mean (average) is the more commonly used for the two estimates, there are situations where the median id to be preferred, as explained in Appendix 4.

(i) Sample average

The sample average is defined as the sum of all the sample values divided by the sample size. Any "Unknown" or "Not Recorded" cases must, of course, be excluded from the calculations and the sample size reduced by the number of cases for which no value is recorded.

In the above systolic blood pressure example, the sum of all the blood pressures equals 4142, and the sample size is 30.

$$\text{Hence, the average} = \frac{4142}{30} = 138.07 \text{ mm Hg}$$

As the calculation is only based on a sample, and blood pressure is difficult to read accurately, the figure should be quoted as 138 mm Hg; 138.07, as calculated, is misleadingly precise.

LTM 410
06082 WA 39

(ii) Sample median

The sample median is defined as a value, such that half of the sample data has values less than it. A simple method for estimating the median is to write out the sample values in ascending order, i.e. from the smallest value increasing to the largest. Re-writing the previously given systolic B.P. values in ascending order in rows of 10, we have :

105	114	117	118	119	120	121	124	125	128
						Q_1			
133	134	134	135	137	138	140	141	142	144
				Median					
146	148	151	151	152	152	153	156	169	195
		Q_3							

The median value is shown by the vertical arrow between the observed (sample) values of 137 and 138; half of the sample blood pressure values, i.e. 15 out of 30, are less than or equal to this value. (At this stage, ignore the Q_1 and Q_3 that are also shown amongst the ascending values). Because the median falls between two readings (sample values), the median would, as a rule, be taken as the mean of the two values,

$$\text{i.e. Median Systolic B.P.} = \frac{137 + 138}{2} = 137.5$$

However, because of the difficulty of recording blood pressure very accurately, it may be better to quote the B.P. to the nearest whole number (integer); in the example, either 137 or 138 mm Hg is close enough.

For large samples, writing out all the data in ascending order is very tedious. An estimate of the median, for large samples, is easily obtained from the percentage frequency table as is explained in Appendix 3.

(iii) Sample proportions and percentages

Proportions and percentages are used to answer such questions as "what part" or "what fraction" of the whole has a certain characteristic. For instance, what proportion of systolic B.P. exceeds the upper limit beyond which most doctors become concerned about the patient's health ? If the upper limit is taken as 150 for illustration, then, in the 30 values given previously, eight cases exceed the limit value and the proportion exceeding 150 is 8/30 = 0.27 approximately.

The percentage equals the proportion multiplied by 100; hence the percentage of respondents exceeding a B.P. of 150 in the sample is : 0.27 x 100 or 27 per cent, usually written as 27%.

The figure above, or below, which certain actions or precautions are initiated (started) is often referred to as the cut-off point or decision value. In many surveys it is of interest to see what percentage of the survey study units either exceed, or fall below, the cut-off value. For instance, for a healthy diet it is often considered that some meat or fish should be eaten at least once a week. The percentage of families who have meat or fish less than once a week, i.e. the percentage of families below this nutritional cut-off point, is of medical and social importance.

(iv) Sample ratios

Ratios are used to measure, or express, the relative size of components (parts) of the sample, or of the population, to each other. The sex ratio, the number of females divided by the number of males in a community, is a frequently quoted ratio and is calculated as :

$$\text{Sex ratio} = \frac{\text{No. of females}}{\text{No. of males}}$$

If the number of females is greater than the number of males, then the sex ratio is larger than one.

The difference between a ratio and a proportion must be emphasised. The above sex ratio has a value greater than one, whereas the proportion of women in the community must be less than one because a proportion measures a fraction of the whole and can therefore never be greater than one.

As an example, consider the population census figures for a Scottish region published in 1985. The regional data given are :

Total males (all ages)	= a	= 357,177
Total females (all ages)	= b	= 388,052
Total population (all ages)	= a+b	= 745,229

Hence, for the region we have :

1) the regional sex ratio :

$$\frac{b}{a} = \frac{388{,}052}{357{,}177} = 1.1 \text{ approximately.}$$

2) the regional proportion of females :

$$\frac{b}{a+b} = \frac{388{,}052}{745{,}229} = 0.5207 = 0.52 \text{ approximately.}$$

(v) Estimated community totals

A knowledge of the totals is often essential if the needs of a community are to be met. When a food shortage exists, a measure of the total amount of food needed is helpful in planning and organising relief supplies. Likewise, the total number of people living in an area must be known if medical services are to be planned and supported in keeping with the community's requirements. A completely accurate knowledge of such totals is not usually necessary as the planning

of resources does not generally demand a high degree of accuracy in the figures used to estimate requirements. Nevertheless, the data on which planning and organisation are based need to reflect the real situation. For example, the planned provision of midwifery services for an estimated total of 10,000 women aged 15 to 45 will not be much out of step with actual needs if in reality there are 10,500, or even 11,000, women in the age group. Yet, if instead of 10,000 there were a total of 20,000, then clearly services only adequate for 10,000 would become greatly stretched, with the result that the quality of health care would decline.

There are several ways of estimating community totals, but two are particularly important in survey work.

They are :

(a) using the sampling fraction

(b) using the sample ratio

(a) Using the sampling fraction

$$\text{Estimated Community Total} = \frac{\text{Sample Frequency}}{\text{Sampling Fraction}^*}$$

Example :

In a survey of a small industrial town of 5983 homes, a random sample of every 20th house, i.e. 299 homes, was to have been selected for enumerating the number of residents and for interviewing. Although a sample fraction of 1 to 20 was initially decided upon, only 294 homes were actually visited. The survey results showed the following age structure :

---

* The Sampling Fraction is defined as the proportion of study units taken into the sample. See Booklet 2 on Sampling.

	Frequency Table of Sample Results		Estimated Community Total	
Age Group	Females	Males	Females	Males
Under 5	34	39	692	794
5 - 14*	67	70	1365	1426
15 - 19	49	51	998	1039
20 - 24	58	59	1181	1202
25 - 29	44	46	896	937
30 - 34	41	40	835	815
35 - 39	39	35	794	713
40 - 49	67	64	1365	1303
50 - 59	63	57	1283	1161
60 +	226	148	4603	3014
Totals :	688	609	14012	12404

In the above example, the sampling fraction actually used is :

$$\frac{294}{5983} = 0.0491,$$ just short of the 1 in 20 originally intended.

The town's estimated total for each age group is then obtained by dividing each of the corresponding survey totals by the sampling fraction. Thus for females aged under 5, the estimated total for the whole town is :

$$\frac{34}{0.0491} = 692.45 = 692$$ to the nearest whole number,

which is the figure shown in the above table. All the other estimated totals are obtained by a similar calculation.

* Age intervals are commonly written in one of two ways, either '5-14' or '$5 < 15$'. The two expressions are to be interpreted in the same way; a child falling into the above age group will have passed his fifth birthday but will not yet have passed his fifteenth.

A Useful Check : Perform the same calculation on the total number of males and females counted in the survey. The answer should be very close to the sum of the estimated totals for all the age groups. For example, as the total number of females of all ages counted in the study was 688, then these computations give :

$$\frac{688}{0.0491} = 14012$$

which is the same as the sum of estimated totals for females for the town. Similarly, for the males of all ages :

$$\frac{609}{0.0491} = 12403$$

which is very close to the sum of the male totals, 12404.

(b) Using sample ratios

Estimated Community Total = Sample Ratio x Total number of units in the community.

Example :

In the survey of the town referred to above, it was found that the number of bicycles owned by the residents of the 294 houses visited was 213. Then the sample ratio of bicycles to houses is :

$$\frac{213}{294} = 0.724 \text{ bicycles per house.}$$

The estimated total number of bicycles owned, by persons resident in the town, is then given by the above ratio multiplied by the total number of houses :

$$0.724 \times 5983 = 4332 \text{ bicycles.}$$

Note : In such a calculation it is **essential** that the unit used in the denominator * of the ratio, in this example a house, is a unit for which the total in the whole community is known. If the total number of these units in the community is not known, then this method cannot be used.

In the same survey of 294 houses, it was required to estimate, amongst those aged 60 and over, the extent to which their medical needs were not being met. Within this age group, the survey found 148 males of whom 26, on examination, required more medical attention than they were receiving. This was mostly due to the lack of initiative on the part of the elderly, fear of seeing a doctor or inability to cope adequately with daily events. Amongst the 226 survey women aged 60 and over, the number requiring additional medical attention was found to be 53. What then is the estimated number of persons aged 60 and over in this town who require more medical attention ?

An approximate method for estimating the additional care required, is to use the ratio of the medical need still to be met for each of the sexes and then to multiply the ratios by the estimated community total for these same age groups, which were estimated in the previous table as 3014 males and 4603 females.

For men, the ratio for the unmet medical needs

$$= \frac{26}{148} = 0.176.$$

Hence, the estimated community total for males

$$= 0.176 \times 3014 = 530.$$

For women, the ratio is $\frac{53}{226} = 0.2345$

and the estimated community total of female persons with unmet medical needs = $0.2345 \times 4603 = 1079$.

---

* In a fraction, the divisor is called the denominator. For example, in the fraction $\frac{13}{17}$, 17 is the denominator whilst 13 is the numerator.

The total number of elderly (over 60) for the community with unmet medical needs is therefore approximately : 530 + 1079 = 1609, or 1600 for all practical purposes.

The above method, although frequently used, provides only an imprecise estimate but it is usually sufficiently reliable for planning purposes. The reason for the imprecision is that the procedure uses two separate sample estimates, the ratio and the estimated total of those aged 60 and over, both of which are unlikely to be completely accurate. The two imprecise estimates are then multiplied together to give the estimated community total. Such imprecision is not likely to be crucial. Even if the true value, instead of being about 1600, were as low as 1400 or as high as 1800, the approximate values for the unmet medical needs are sufficient to allow appropriate remedial action to be planned.

2. Estimation of Variability

Variability is concerned with the extent to which the variables between study units differ one from the other. A variable does not only have a typical value, such as the mean or median, but also has variability, because few, perhaps none, of the sample results (values) are the same. Variability is a vital concept (idea) in statistics and survey work.
There are several ways of expressing, or measuring, the variability of data. The most commonly used are :

(a) the range

(b) the quartiles

(c) the standard deviation

of which the standard deviation is the most useful index of variability, but it is also more complex.

(i) Sample range

The range consists of just two values, the lowest and highest values in the sample. Thus for the above sample of the systolic B.P. of 30 men, the range is 105 to 195. The range is usually quoted (shown) alongside the mean or the median so that the readers of the report have some idea not only of the typical value given by the mean or the median, but can also see how widely the data fluctuates without having to look at all the results.

Although the range is useful and informative, it should not be used as the only measure of variability. The main objection to the range is that it makes use of only two values, the lowest and highest in the sample. The extreme values do not indicate whether the lowest and the highest are close to or far away from the other results, a fact that is of importance in reporting medical data, where pathological (abnormal) results, can be very different and far away from normal and healthy results. Moreover, the range tends to widen as the sample size increases, because as the sample becomes larger, there is a greater chance that a value smaller than the previous smallest value will come into the study. Similarly, the previous largest value may be exceeded as the sample size increases.

(ii) Quartiles

There are three quartile points (values); these divide the sample distribution into four parts, or quarters.

The first quartile, Q_1, is a value such that a quarter of all the sample results are less than or equal to the value of Q_1.

The second quartile, Q_2, is a value such that half (two quarters) of all the sample results are less than or equal to the value of Q_2. The value of Q_2 is the same as that of the median, which was defined earlier.

The third quartile, Q_3, is a value such that three quarters of all the sample results are less than, or equal to, the value of Q_3.

An easy way of finding the quartile points, Q_1, Q_2, Q_3 is to arrange the sample values (results) in ascending order and then to insert Q_1 at the point where a quarter of the sample values are less than or equal to that value. Similarly for Q_2 and Q_3.

An illustration of the method is given by the above 30 systolic blood pressure results,* written in ascending order; we see that

$$Q_1 = \frac{121 + 124}{2} = 122.5 \text{ and } Q_3 = \frac{151 + 151}{2} = 151;$$

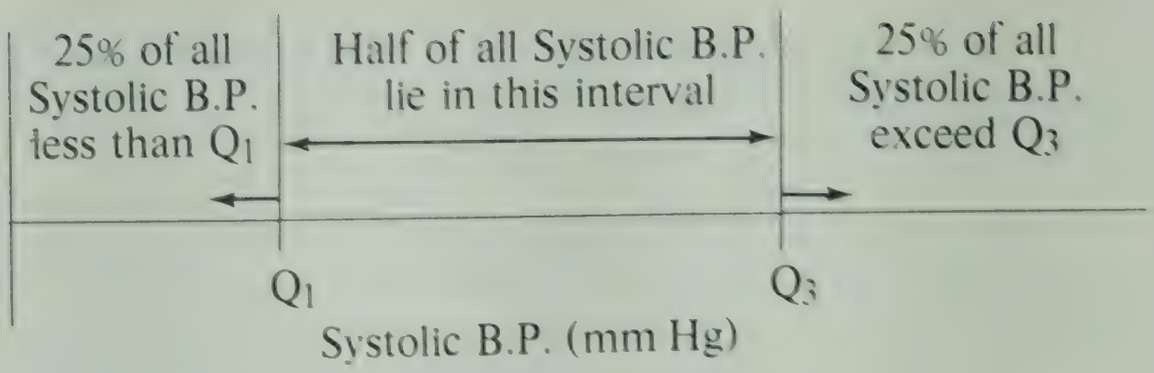
$$Q_2 = \text{Median} = 137.5.$$

As before, if the quartiles fall between two values, the mid-point of the adjacent values is taken to be the approximate quartile figure.

The importance of the quartiles lies in the interval Q_1 to Q_3. The interval from Q_1 to Q_3 is the sample estimate of an interval within which half (50%) of the population values lie. Another way of stating this, is to realise that Q_1 is an estimated value such that one quarter (25%) of the population values are less than Q_1; similarly, Q_3 is a sample estimate of a value such that only one quarter (25%) of the population values exceed it, which is, of course, the same as saying three quarters of the results are less than Q_3. Yet another way of looking at the meaning of Q_1 and Q_3 is to think of a variable, such as systolic pressure, measured along a line :

---

* See page 42 and 52.



(iii) Standard deviation

The standard deviation* and some of its applications are briefly discussed in Appendix 5.

3. Verification and Checks

Verification (checking) at all stages of the survey analysis is absolutely essential.

Verification must be done at each stage **before** the next stage starts. Coding has to be checked before the questionnaires are sorted into groups. When the sorting has been checked, and found to be correct, then extraction and summary sheets have to be carefully verified.

In survey work the best, and usually the only, convincing way of checking is to have the whole job done a second time; the second results are then compared with the first set. Moreover, as far as possible, the person doing the checking should not be the same person who did the coding or extraction the first time.

Apart from checking in this way, i.e. by doing the work independently a second time, there should also be spot checking and monitoring of the ongoing work by the survey organiser.

---

* Most introductory books on statistics will discuss the standard deviation, how it is calculated and used.

Remember, a mistake made during coding, but not discovered, means the error is carried forward into data extraction, into the computations and finally into the report itself; if the error is sufficiently serious it could affect the conclusions reached. There are three golden rules :

1) Never proceed to the next stage of survey analysis until the current stage has been independently and carefully checked.
2) Where, for purposes of checking, a procedure is repeated a second time, the comparison of the two sets of results **must,** in addition, be checked by the organiser as well.
3) Whenever arithmetic checks are available, as is the case with contingency tables, they must be carried out as well.

4. Simple Graphical Presentation

A graph or diagram, if properly drawn without too much detail, provides an easily understood picture of the data. A suitable diagram is easier to grasp and leaves a more permanent impression of the main features of the data than do arithmetical and statistical procedures.

There exist a great number of graphical methods and ways of presenting graphs to meet special needs. However, four types of diagram are commonly used and are particularly useful for presenting survey data or for showing the connection (association) between two variables.

The aim of every diagram should be to convey **essential** information in a simple and direct manner, and this can be achieved by following the guidelines given below :

(a) A diagram should only show essential features; excessive detail destroys its clarity and simplicity.

(b) A diagram needs to have a clear, descriptive title which includes the date and place of the study; sometimes a short description of the data is added.

(c) The size of the diagram must neither be too small nor excessively large, as either will detract from its clarity. As a general rule, a diagram should be between 5 and 15 centimetres in length and height.

(d) Where appropriate, the use of different colours, different shadings and differently drawn lines increases the contrast between different areas and lines in the diagram. In this way, even a fairly complex graph can be made considerably easier to read and to understand.

(e) To have maximum effect, a diagram should be neatly and carefully drawn.

(f) A diagram should state, either on the diagram itself or in an accompanying description, the total number of cases (observations) on which it is based.

(i) Pie chart

The pie chart consists of a circle divided into segments (wedges); the size (area) of a segment is proportional to the percentage of cases belonging to the group it represents.

The table and our pie charts below refer to a survey of alcohol drinking patterns in Scotland *. The drinking of alcohol, particularly if done to excess, creates many social and medical problems. Because of these problems, and the absence of reliable information, a survey was undertaken to study the drinking habits of the population. One of the aims was to establish the different drinking habits and preferences between men and women and the two principal social groupings, called "manual" and "non-manual" (for type of employment). Whilst the table below expresses the differences in percentages, the pie charts more graphically display the difference in the size (area) of the pie chart wedges. To help distinguish between the pie chart segments, they are hatched or shaded in different ways. A key is provided showing that beer drinkers are represented in the pie chart by the black area while the other segments are distinguished by various shadings and hatchings (lines drawn diagonally).

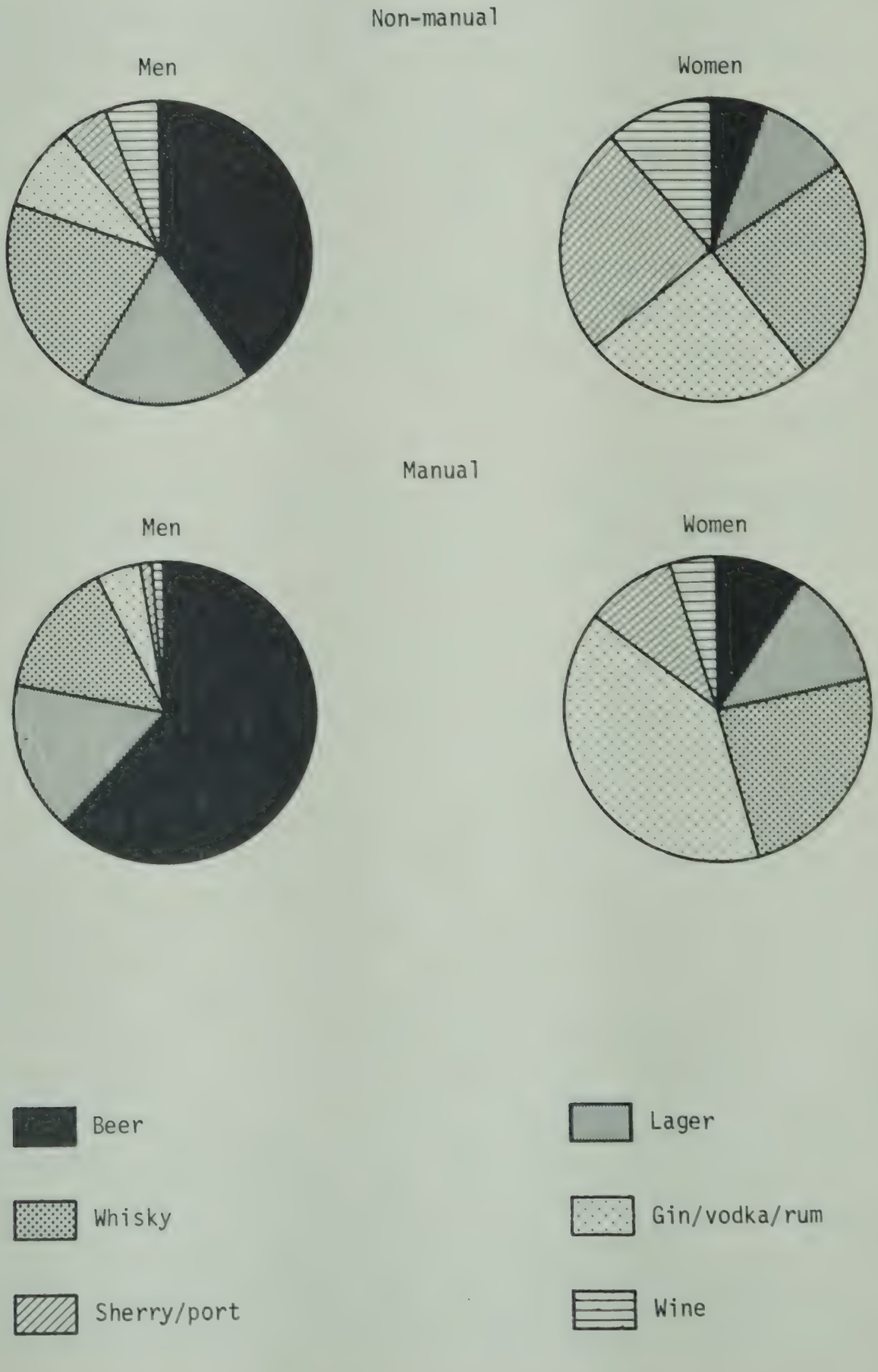
---

* Susan E. Dight, Scottish Drinking Habits, Office for Population Census and Statistics, Her Majesty's Stationery Office, 1976.

Regular drinkers, Scotland 1972: Percentage of total amount, by sex and social class of head of household

Alcoholic beverage	Male regular drinkers		Female regular drinkers	
	Non-manual	Manual	Non-manual	Manual
	%	%	%	%
Beer	40.3	62.0	6.0	9.5
Lager	18.3	16.2	9.7	12.2
Whisky	21.8	14.8	23.9	24.2
Gin, vodka, rum	9.2	4.7	24.5	39.6
Sherry/port	4.9	1.4	24.4	9.8
Wine	5.5	0.9	11.5	4.7
All regular drinkers	100.0	100.0	100.0	100.0

Regular Drinkers, Scotland 1972 :
Mean proportions of each beverage consumed by men and women in the non-manual and manual classes.



The method of drawing a pie chart can be demonstrated using data from an Indian * survey concerning breast feeding. Some questions were asked about the current employment of the head of household, because it was thought the size of family and breast feeding were influenced by economic and educational factors. The survey findings were as follows :

Head of Household Type of Employment	Frequency	Percentage
1) Not employed	26	3.3
2) Daily wage earner	177	22.2
3) Casual worker	53	6.6
4) Part-time regular employee	13	1.6
5) Full-time regular employee	522	65.3
6) Other	8	1.0
	799	100.0

Step I :

Condense the table into a few large groups, remembering that very small percentages will not show clearly on the diagram. The larger groups so formed must, of course, remain meaningful, leading, in the above example, to three main classes of employment :

* The Dharavi Project, 1985 : An investigation of infant feeding patterns in the major urban slum of Dharavi, Bombay. Unpublished report.

Head of Household Employment	Frequency	Percentage
Daily wage earner	177	22.2
Full-time regular employee	522	65.3
Part-time employment, unemployed and others	100	12.5
	799	100.0

Step II :

Draw a circle of suitable size. Show the radius; any position will do for the radius, although the 12 o'clock position is commonly used as the starting point, as is done in this example.

Step III :

The angle that belongs to each of the pie chart segments is calculated by ;

3.60 x percentage represented by that segment.

Thus for :

1) Daily Wage earners, the angle = 3.6 x 22.2
= 79.9 degrees

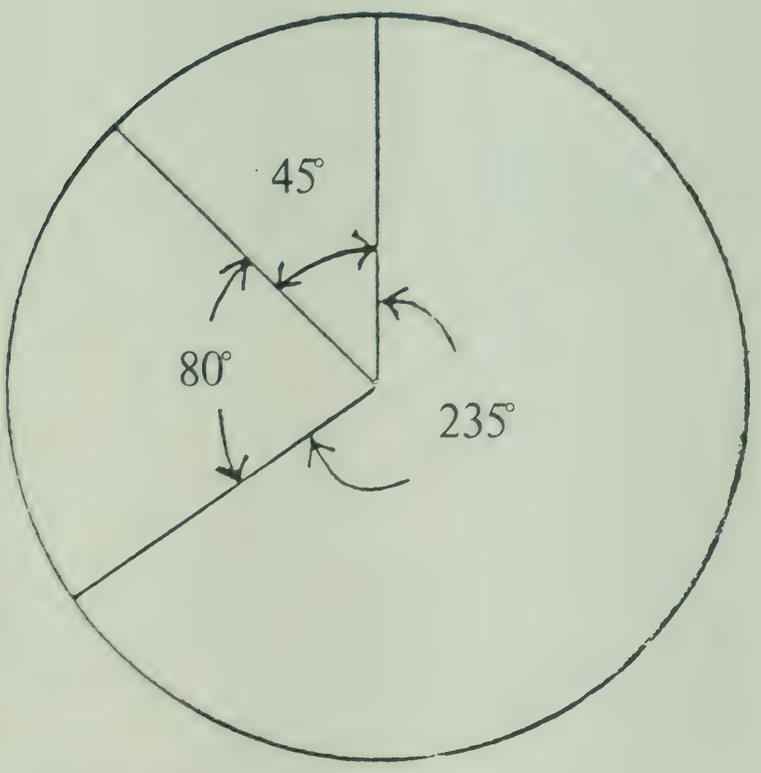
2) Full-time Regular employees, the angle = 3.6 x 65.3
= 235.1 degrees

3) Part-time, etc., the angle = 3.6 x 12.5
= 45.0 degrees

A useful check : The sum of all angles must add up to 360.
Note : For the above, 79.9 + 235.1 + 45.0 = 360.

Step IV :

Draw in the segments, using the angles calculated at Step III to determine their size (area).



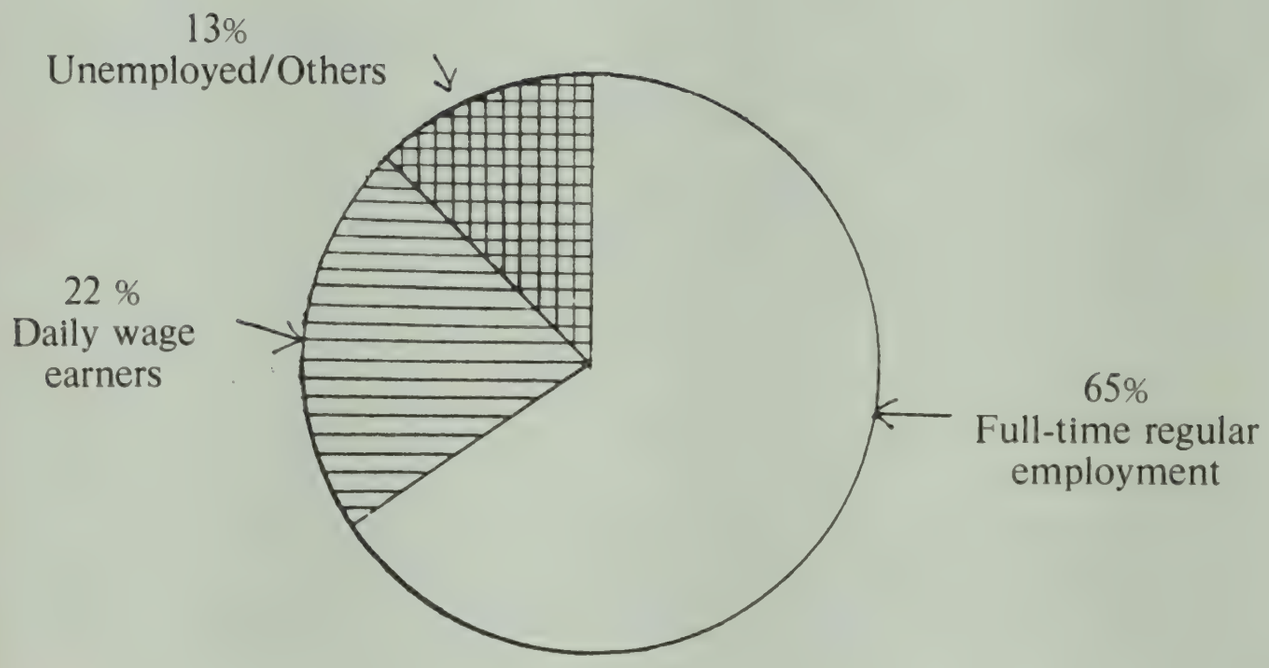
Note : the angles have been rounded to the nearest whole degree as there is no need for greater accuracy in a diagram of this kind.

Step V :

Write in the title, add a description if it is thought useful, state the number of cases and shade or colour the segments as appropriate.

Employment Status of Head of Household, Dharavi, India, 1985.

Sample based on 799 cases



Note :

1) The percentages have been rounded to the nearest whole percent. The percentages calculated from the survey data are estimated values and to quote them to several decimal places suggests a misleading precision.

2) In the example, two of the segments have been shaded, i.e. diagonal lines drawn in. The shading goes in different directions and provides a contrast between the segments. One of the segments has been left blank (unshaded).

3) The description of each segment and its percentage has been written alongside the segment. The number of cases on which the data is based is stated beneath the title.

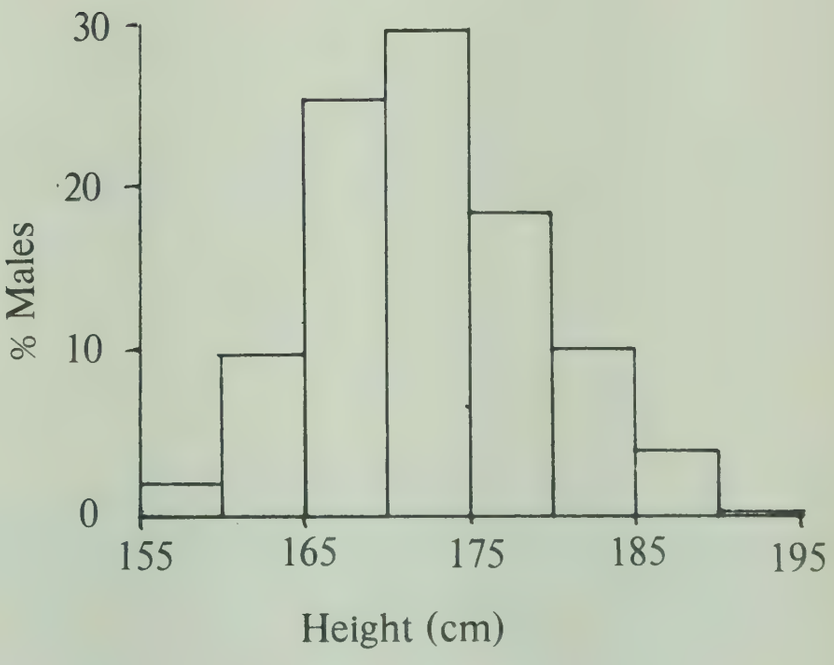
(ii) Histogram

The histogram is a pictorial representation of a table which can be either a frequency or percentage table. The latter is more usual as it is easier to compare tables of similar data if expressed as percentages.

The histogram consists of a series of adjacent rectangles, with the class interval taken as the base (bottom) of the rectangle; the area of the rectangle is proportional to the percentage (or frenquency) that it represents. An example is provided by the distribution of height of a random sample of 448 men aged 45-54.*

Height (cm)	Freq.	%
155-159	9	2.0
160-164	44	9.8
165-169	114	25.4
170-174	133	29.7
175-179	83	18.5
180-184	46	10.3
185-189	18	4.0
190-194	1	0.2
	448	100.0

Heights of a Random Sample of 448 Edinburgh-Fife Males (Age 45-54) 1980



* Edinburgh-Fife Heart Study (1980).

The histogram, like all statistical diagrams, must have a clear title and a description explaining the type of data represented. The scale used for the histogram should be shown as well as the number of cases on which the histogram is based. The size of the histogram for most purposes should be between 5 and 15 cm in both directions, i.e. for both its height and its base.

Special care must be taken when choosing the scale; there are, in fact, **two** scales that have to be decided :

(a) the scale to be used for the class intervals

(b) the scale for representing the **area** of each of the rectangles. In the usual case, where all the class intervals are of equal length, it is sufficient, and easier, just to choose a scale for the height of the rectangles, as was done for the above histogram.

The steps for drawing the histogram, assuming all the class intervals are of equal length, can be illustrated using the example previously given for 30 diastolic blood pressures.*

78	72	82	84	70	85
116	87	69	75	81	87
76	78	82	85	84	74
77	85	76	94	101	88
87	89	98	99	77	81

A sample size of 30 is rather small for calculating percentages, so the example will show the frequency histogram. The same steps apply to the construction of the percentage histogram, although the scale may then need to be changed.

---

* Selected for purposes of illustration from the total of 448 men in the Edinburgh-Fife Heart Study.

Step I :

Choose a suitable scale for the base so that the length for the whole range of the variable, using the chosen scale, is somewhere between 5 to 15 cm In the above example of diastolic blood pressures, the sample values range from 69 to 116, i.e. a range of 47. A scale of 5 mm Hg per cm will therefore give a suitable base length of 11 cm if the first interval starts at 65 mm Hg and the last interval ends at 120 mm Hg. Next, if this has not already been done, set out the frequency table, using equal class intervals and using the scale decided upon.

Step II :

Choose a suitable vertical scale. In the example given below, the maximum frequency is 8, hence a scale of 1 cm = 1 observation will give a height of 8 cm for the largest of the rectangles.*

Step III :

Draw the histogram using the scales chosen in Steps I and II. The resulting histogram is shown below, next to its frequency table.*

Step IV :

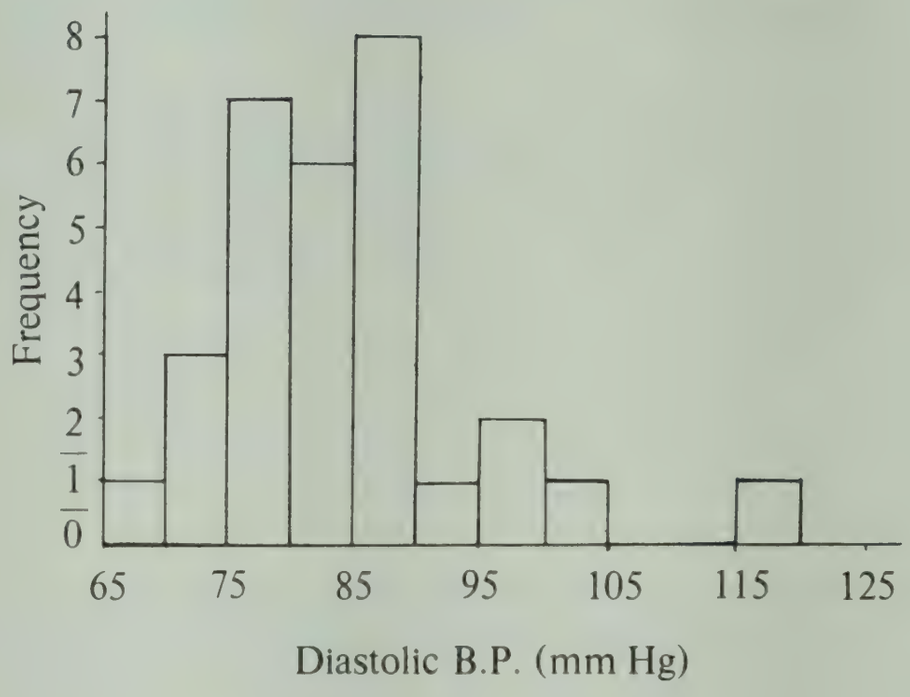
Insert a title, legend (description) and the scale on the diagram; the total sample size used for the histogram should be indicated.

---

* The histogram shown along side the frequency table was originally drawn to this scale, but it has, for the purpose of printing, been reduced in size.

Diast. B.P. (mm Hg)	Freq.
65-69	1
70-74	3
75-79	7
80-84	6
85-89	8
90-94	1
95-99	2
100-104	1
105-109	0
110-114	0
115-119	1
	30

Distribution of Diastolic B.P.
30 Edinburgh-Fife Males (Age 45-54)
1980



(iii) Scatter diagram

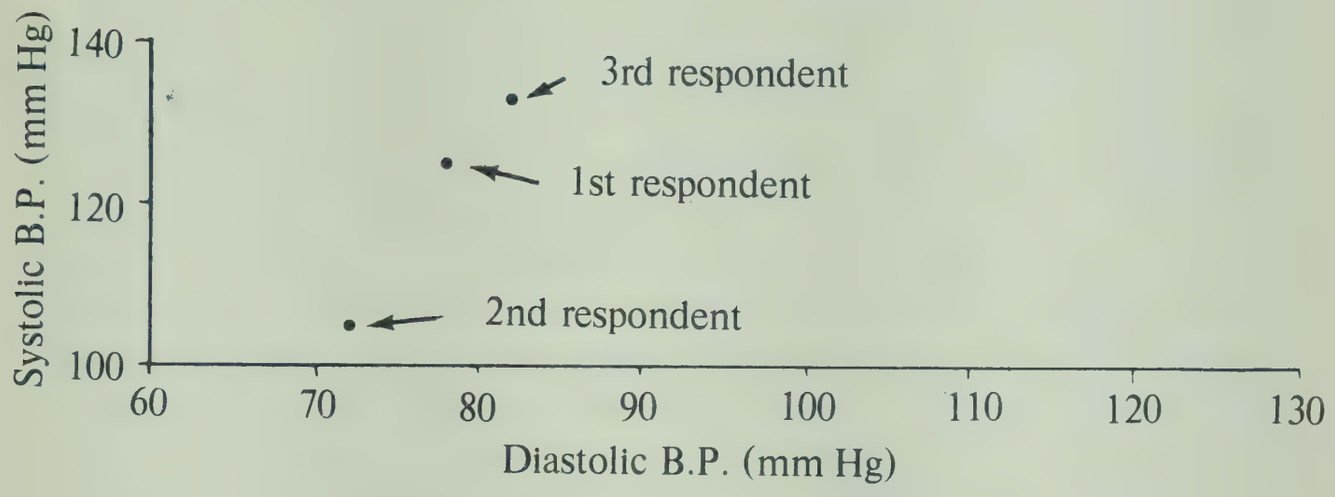
The principal use of the scatter diagram is to study the relationship between two variables. Typical examples are the relationship between :

(a) sytolic and diastolic pressures

(b) height and weight.

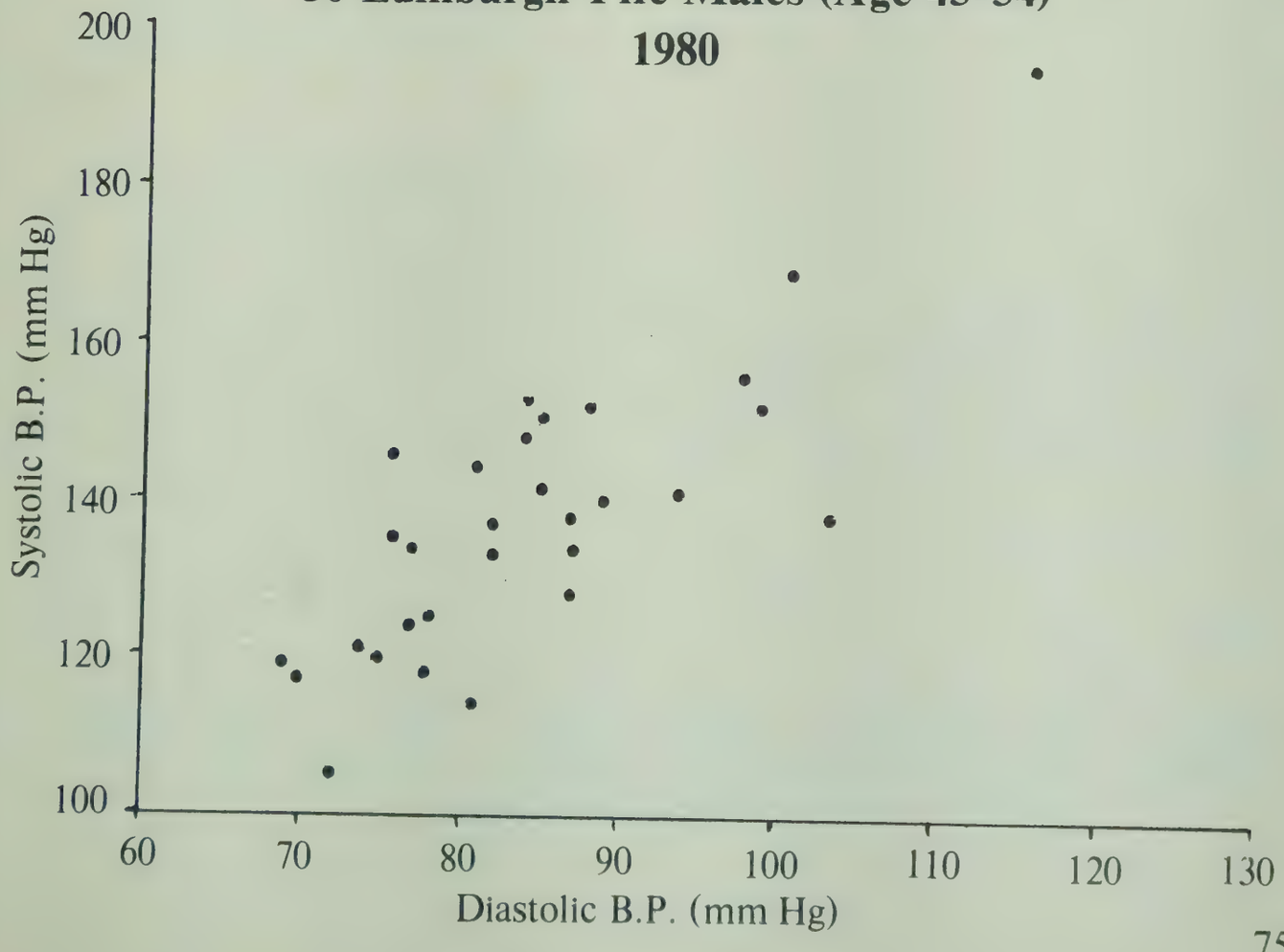
Examples of just such data are provided by the Edinburgh-Fife Heart Study given earlier (p. 42). A suitable horizontal (across) and vertical (upwards) scale for the variables is chosen in the usual way. Plot the values of the two variables on the graph; note that this gives a **single** point for each respondent. The plot of the first three respondents is shown below :

	Systolic	Diastolic
1st respondent	125	78
2nd respondent	105	72
3rd respondent	133	82



Proceeding in the given way for all the 30 cases, we obtain the following scatter diagram.

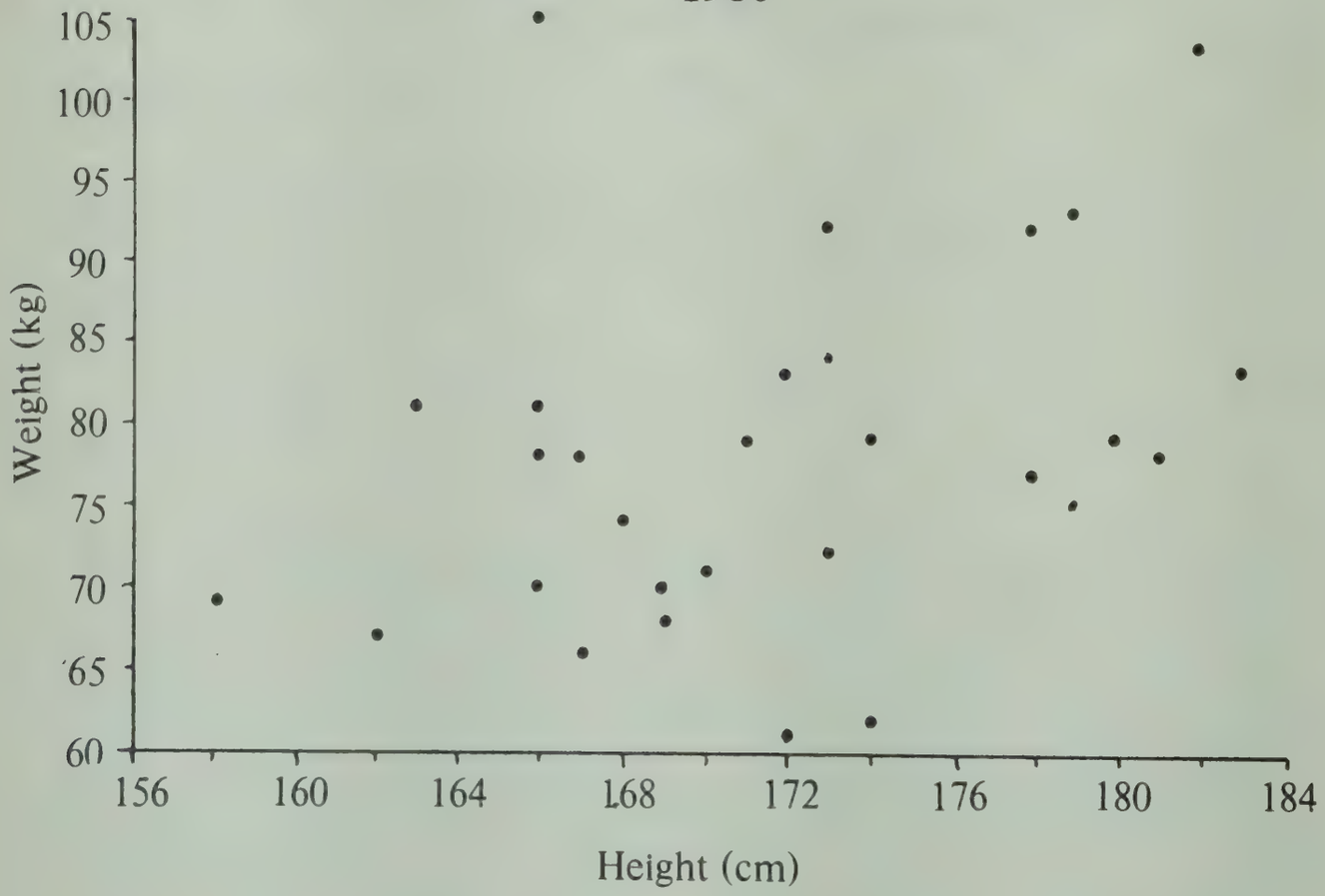
Systolic B.P. against Diastolic B.P.
30 Edinburgh-Fife Males (Age 45-54)
1980



With most respondents, there exists a fairly close relationship (association) between their systolic and diastolic pressures, as can be seen from the diagram. The points on the scatter diagram, because of the close relationship, are reasonably close together and show a clear positive trend, i.e. the chances are that the respondents having a high diastolic pressure will also have a raised sytolic reading.

Where the relationship between two variables is less well defined, i.e. the relationship between them is not very close, the scatter diagram will have its points spread more widely. An example is provided by the same 30 cases from the Edinburgh-Fife Heart Study, but this time using height and weight. Height and weight are also positively associated, as one would expect. A taller person will usually be heavier than a shorter person, but the relationship is not so well defined as can be seen in the greater spread of points in the scatter diagram below :

Weight against Height
30 Edinburgh–Fife Males (Age 45-54)
1980



As with all diagrams, the scatter diagram should have a descriptive title, a suitable scale clearly shown and the number of cases included in the graph should be stated. All the requirements are clearly satisfied in the above two examples.

Unfortunately, the scatter diagram becomes too cluttered (full) when the number of cases is much above 50 or 60; the graph then loses its clarity and the relationship between the variables becomes blurred. The method described below, under Trend diagrams, is suitable for large samples.

(iv) Trend diagrams

As the name suggests, trend diagrams are a graphical means of studying the trend, i.e. the way a particular measurement changes as some other variable increases. The hospital fever chart is a common example of a trend diagram; it traces out how the patient's temperature changes with time.

Such charts, in which one of the variables is time, measured in either hours, days or years, are often called time charts or time series.

(a) Time charts

Almost without exception, time charts show time running horizontally, i.e. across the page from left to right, whilst the other variable has a vertical (up and down) scale. The following typical example is taken from the British Medical Journal.*

---

* Seasonal variation and time trend of death from asthma in England and Wales 1960-82. A. Khot and R. Burns, Vol. 289, 1984.

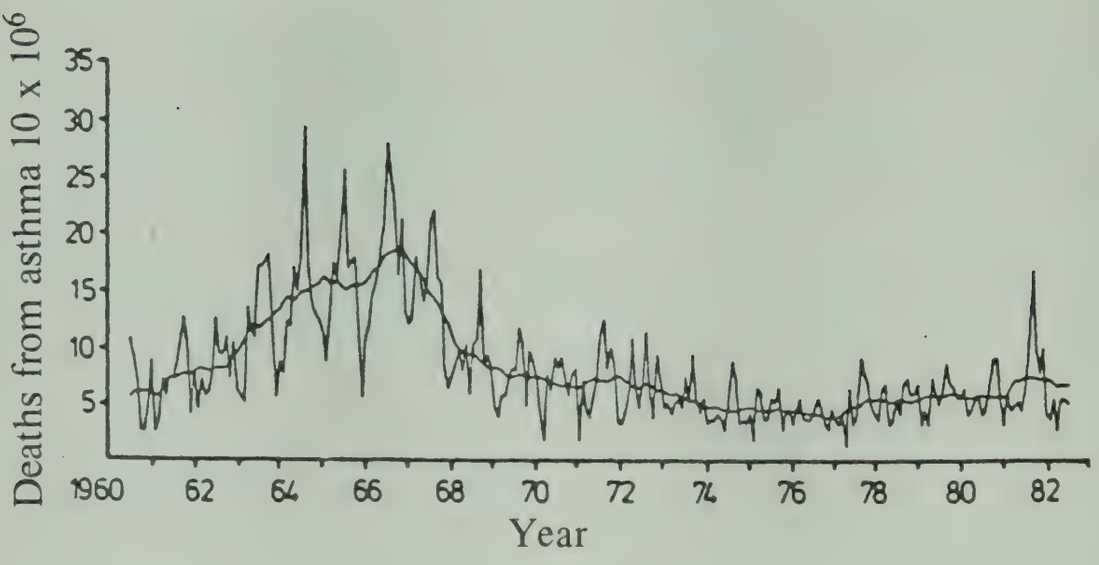


FIG 1 - Monthly mortality rates for asthma (age 5-34), with superimposed trend, in England and Wales 1960-82 (source : Office of Population Censuses and Surveys).

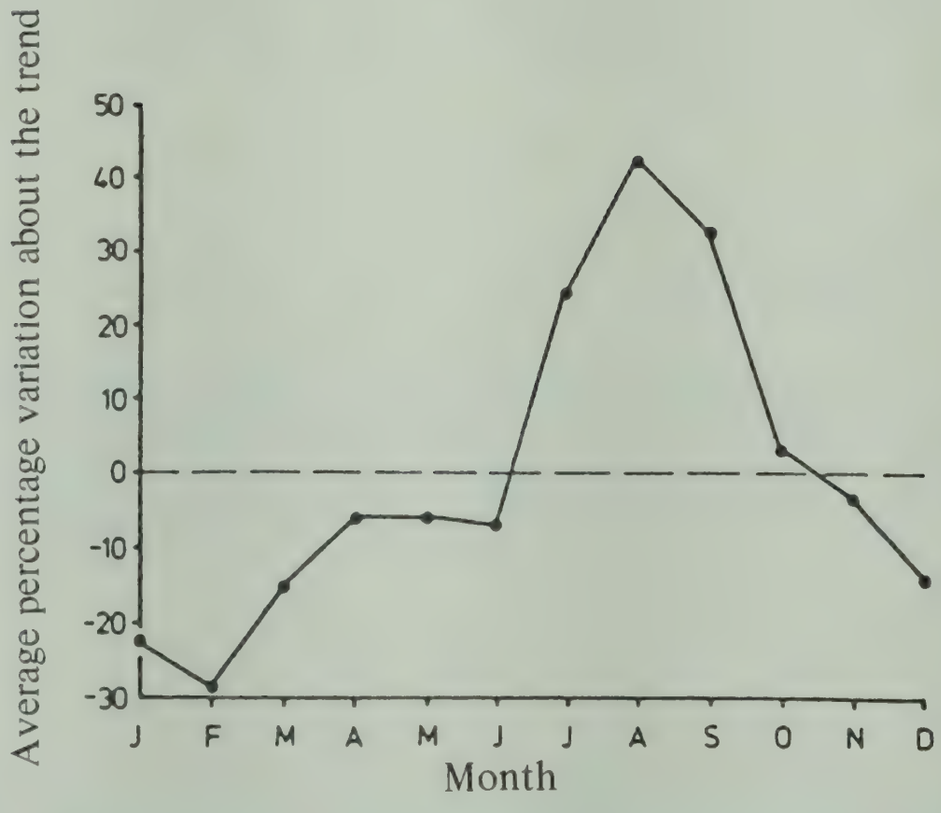


FIG 2 - Average monthly variation in deaths from asthma in 5-34 year age group in England and Wales 1960-82.

The two time charts relate to the same data over a 23 year period, but each emphasises a different aspect of the information.

Figure 1 shows the ratio (number of deaths per 10 million population in the age group 5 to 34 years) plotted monthly against time. Careful examination of the plotted data reveals two medically important features :

(a) the trend of deaths from asthma was rising between 1960 and 1967 and thereafter, there was a decline. First there was a rapid decline between 1967 and 1970, followed by a levelling out in later years to about 5 deaths per 10 million per month. The decline is emphasised by the superimposed trend line.

(b) within **each** year, the ratio of deaths per 10 million peaks at certain times of the year. This suggests that climatic conditions are particularly adverse for asthma sufferers during some months and less so at other times. However, this aspect of the data is not particularly well shown by Figure 1.

Figure 2 is designed to show the typical monthly variation, January to December, in the death rate (per 10 million). To do this the authors proceeded as follows :

1. Starting with the first year, 1960, the average monthly death rate per 10 million was calculated for the whole of that year. This average was then subtracted from the twelve individual monthly rates for January to December 1960. Where a monthly rate was below the average rate for the year, a negative difference resulted.

2. Each of these twelve differences, January to December, 1960, was next expressed as a percentage variation from the average for 1960. A negative difference, as calculated under (1), yielded a negative percentage variation.

3. This procedure, as outlined under (1) and (2) above, was done for each of the 23 years, 1960 to 1982 inclusive. Thus there were 23 percentage variation results for each of the twelve months, spanning the period 1960 to 1982. For each month, an average percentage variation was calculated; for instance, the January average percentage variation was obtained by summing the 23 individual January percentage variations, as obtained under (2), and then dividing the sum by 23. Similar average percentage variations were calculated for each of the other months.
4. The average percentage variation, as obtained under (3), was then plotted against the corresponding month, as shown in Figure 2. The dotted horizontal line (at zero) represents the theoretical line we would expect if there was no monthly variation in the death rate per 10 million.

The diagram shows extremely well that, for England and Wales, the months of July to October, i.e. late summer and early autumn, are the worst for asthma sufferers.

The two diagrams drive home three important lessons :

(a) the illustrative power of trend diagrams.
(b) more than one diagram may be required to emphasise different aspects of complex data.
(c) the trend diagram allows considerable flexibility in designing diagrams to illustrate important aspects of the data.

As with other types of diagrams, care must be taken to provide a meaningful title and, where appropriate, a short description of the data. The scale and units of measurements must be clearly indicated, as must the number of cases, where it is not self evident from the text.

(b) Other applications of the trend diagram

Trend diagrams have a wide range of applications and provide the simplest method for studying the relationship between two variables, especially where the sample size is large.

By convention, the horizontal axis (across the page) is used for the so called independent variable, i.e. the measurement that is considered not to be affected, or is less affected, by the other measurement being studied. For instance, to examine the relationship in adults between height and weight, height would normally be taken as the independent variable as weight does not have much effect on height, whereas weight is substantially affected by height; a taller person will, as a rule, weigh more. Hence height is usually shown along the horizontal * axis.

In the previous discussion of the scatter diagram, the results of the first 30 men from the Edinburgh-Fife Heart Study were given. The study actually examined 448 men aged 45-54, but this number is too large to be plotted as a scatter diagram. The relationship between diastolic and systolic blood pressure, for all 448 men, can however be studied by two other simple methods :

(i) the contingency table

(ii) the trend chart.

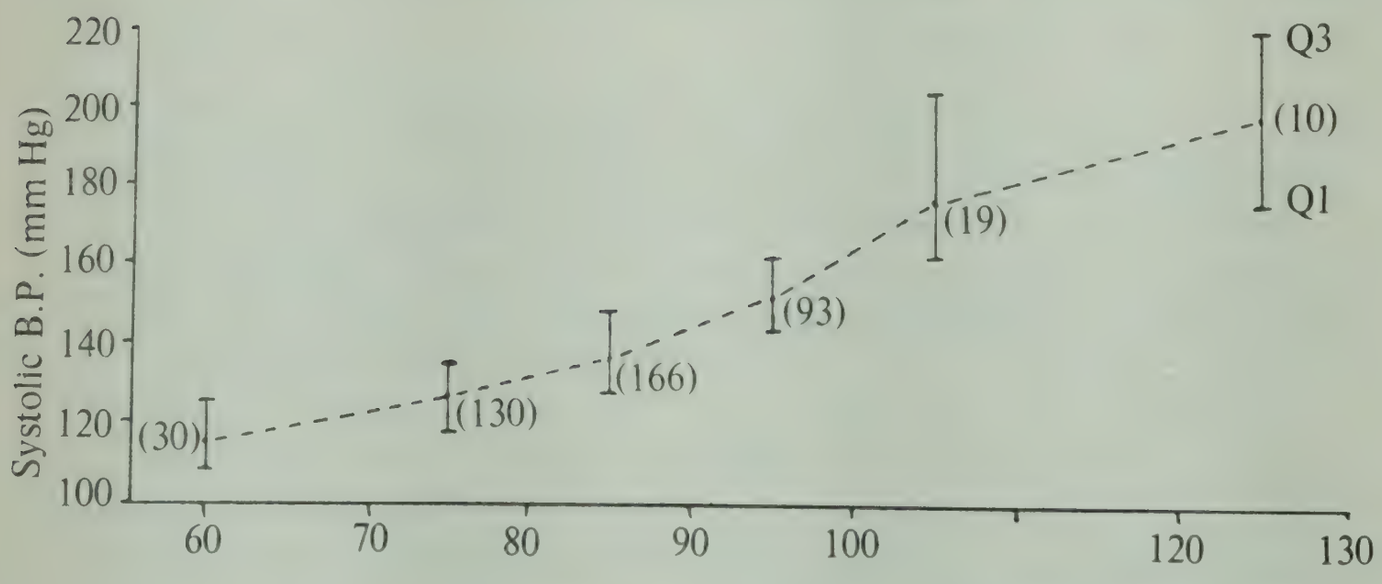
The trend chart is most easily constructed from the contingency table which is shown below. The association between systolic and diastolic pressures is demonstrated by the largest cell frequencies being located in the **diagonal** cells from top left to bottom right, showing that as diastolic pressure increases so does systolic for most of the 448 respondents. However, the effect, as revealed in the table, is not obvious at first glance. The table has to be carefully studied before the relationship becomes clear. In contrast, the trend chart will reveal the association immediately and without ambiguity as is seen below.

---

* In paediatric growth tables, this would be reversed : age, which is a measurement of time, is always shown along the horizontal axis. Time is considered to be the independent variable which has an effect on growth.

Systolic B.P. (mm Hg)	Diastolic B.P. (mm Hg)						Row Totals
	50-69	70-79	80-89	90-99	100-109	110-139	
100-119	20	37	5	0	0	0	62
120-139	9	80	96	1	0	0	201
140-159	0	12	59	51	4	1	127
160-179	1	0	6	24	7	2	40
180-239	0	1	0	2	8	7	18
Column Totals	30	130	166	93	19	10	448

Median Systolic B.P. against Diastolic B.P.
448 Edinburgh-Fife Males (Age 45-54) 1980



The following points are worth noting about the above trend diagram :

1. The information plotted is partially derived from the contingency table for systolic and diastolic pressures. Within each diastolic class interval, the systolic median * and quartiles * are calculated.

2. For each diastolic class interval, the systolic median is plotted against the **mid-point** ** of that diastolic class interval. The corresponding quartiles, Q_1 and Q_3, are also plotted and joined by a vertical line to indicate on the diagram the interval within which 50% (half) of the sample systolic pressures lie.

3. A dotted line is drawn between adjacent median points to emphasise the trend and the relationship between systolic and diastolic pressures.

4. The systolic intervals with large frequencies are more reliable than those based on smaller numbers. The number of cases for each interval is therefore shown in brackets to indicate on which points most reliance can be placed.

5. The quartiles, Q_1 and Q_3, do not always lie symmetrically about the median, a fact that is particularly noticeable with the fifth interval which contains 19 cases; the median does not always lie half way between Q_1 and Q_3.

---

* The method of estimating the median and quartiles from frequency tables is shown in Appendix 3. An alternative, favoured by some statisticians, is to plot the mean of the systolic pressure for each column, instead of the median; if this is done then Q_1 and Q_3 are also replaced by some other values. This alternate method is explained in Appendix 5.

** The length of a class interval is taken from the start of the interval to the beginning of the next. Thus in the above table,the Diastolic B.P. interval, 50-69, has a length of 20; its mid point is therefore 60.

(c) "How" and "What" to plot

The above trend diagram of systolic against diastolic pressures is instructive. Diagrams presenting similar problems frequently occur in survey work.

The stepwise procedure, to obtain the values to be plotted, is as follows :

1. Choose the variable best suited for plotting along the horizontal axis and divide it into appropriate class intervals. It is an advantage if the intervals are of equal length unless there are good reasons for preferring unequal intervals.
2. Separately, for each of the above class intervals, write down all the values of the second variable falling within each of the first variable intervals.
3. Calculate, for **each** of the first variable intervals, an appropriate statistical index. Often the index is the average or the median for those cases falling within a particular interval.
4. Plot the index calculated under Step 3, against the midpoint of the corresponding (horizontal) class interval.

The four steps are illustrated below, using the first 30 respondents from the Edinburgh-Fife Heart Study.

Original (Raw) Data

Height	173	180	169	**158**	179	169	162	178	167	168
Weight	72	79	68	69	93	70	67	77	66	74
Height	166	168	182	170	179	167	178	173	166	171
Weight	105	74	103	71	75	78	92	92	70	79
Height	166	174	172	163	172	174	173	166	181	**183**
Weight	78	79	83	81	61	62	84	81	78	83

The range for heights is 158 to 183 cm; hence six intervals of 5 cm each would normally be appropriate; however, the first two will be combined because they contain less than three weights each.

Step I and II :

Class Intervals Height (cm)	Individual Weights (kg)	Total Weight (kg)
155-164	69, 67, 81	217
165-169	68, 70, 66, 74, 105, 74, 78, 70, 78, 81	764
170-174	72, 71, 92, 79, 79, 83, 61, 62, 84	683
175-179	93, 77, 75, 92,	337
180-184	79, 103, 78, 83	343

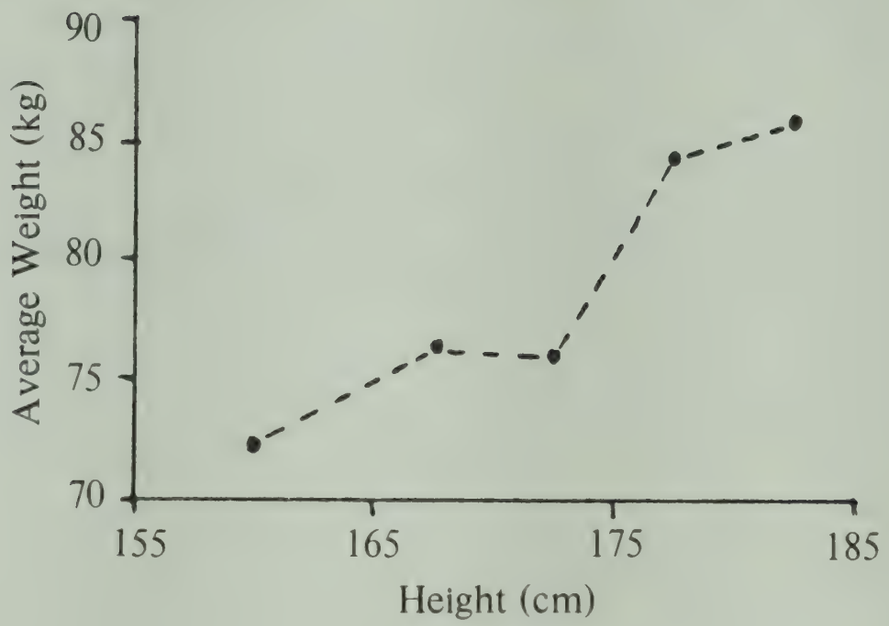
Step 3

The mid-points of the class intervals and the corresponding average weight are calculated, giving :

Class interval Mid-Point (Height)	160.0	167.5	172.5	177.5	182.5
Average weight within the class interval	72.3	76.4	75.9	84.3	85.8

Step IV :

Average Weight against Height
30 Edinburgh-Fife Males (Age 45-54)
1980



In the above outline of "How" to plot a trend diagram, the instruction under step III is to calculate an appropriate statistical index. In most survey applications, the appropriate index is a rate, a percentage, an average or a median. The choice is left to the survey organiser as it depends entirely on the nature of the data. Difficulty is sometimes experienced in choosing between the average and the median; this is discussed in Appendix 4.

Occasionally, it is desirable to plot not only the actual median point but also an interval within which a known percentage of the sample falls. If the median is chosen for plotting, then the interval points should be the quartiles, Q_1 and Q_3.*

The method for estimating the median and quartiles for small samples has already been explained (see p. 60). The method is tedious for large samples and a procedure suitable for large samples is oulined in Appendix 3.

---

* What to do if the mean is chosen is discussed in Appendix 5.

Section D : Report Writing

1. Types of Report and their Dissemination

After completing the analysis of the survey results, the information obtained, together with the conclusions reached, needs to be written up as a permanent record and to provide a means for disseminating (communicating) the information. Various types of reports and summaries can be written depending upon the survey objectives and the readership the organiser has in mind. The following should be considered :

(a) A full and detailed report to serve as a permanent record and reference manual (book) for use by the organiser and others directly concerned with the community's problems and services.

(b) A report for distributing amongst those who provide similar services or face similar problems in their own community, or who, for other reasons, are interested in the problems and situations studied.

(c) A report to provide information about, and insight into (understanding of), the communities' services and problems for those who may be in a position to alleviate (improve) the situation or who can provide additional resources.

(d) A short report or summary for the many people who may at some stage have given support to the study or who worked for the study, such as the community leaders and interviewers.

(e) A popularised summary of the most important facts and conclusions of the study for more general distribution in the community and at meetings.

It is sufficient, in most surveys, to use the same full report for both (a) and (b), a reduced (less comprehensive) report for both (c) and (d) and then a final one - or two-page leaflet for wider distribution.

Generally, it is the full report that is written first. The other reports and extracts, being shorter, are then written by selecting from and condensing the material and topics as set out in the main (full) report.

The scientific and professional journals are another means of disseminating the study results. Some of the survey information may turn out to be particularly important, especially if the findings are quite new and unsuspected; in such cases the results can be written up and submitted (sent) for publication to a medical, sociological or other suitable journal. All journals have their own rules as to how articles must be written, how long they may be, the maximum number of tables and graphs permitted, and so on. The guidelines to authors can usually be found in previous issues of the journal or can be obtained by writing to the editors.

In some situations, the findings of a survey may be of general, not just specialist, interest. If it is so, then a newspaper or popular magazine may also be willing to publish something about the study. A newspaper article will, of course, be written in a different style to that used in writing a scientific report. The editors and reporters from the newspaper will usually do the writing after having first discussed the study with the organiser. Great care must be taken that the newspaper reporter and editor really understand the subject and the survey conclusions, otherwise what appears in the newspaper may be very different from what the organiser intended to say.

There also exists the danger that some newspapers which have special political interests, or which present minority views, may deliberately exaggerate, or in some other way distort the conclusions reached in a community study. Such distorted newspaper, or indeed radio and television reporting, can do much harm in that it can give the population a false impression of the purpose of the survey and of its findings.

2. Structure and Content of the Report

The structure of the full report, i.e. the number of sections and the order in which they appear, will vary according to the organiser's aims and objectives as well as the readership for whom the report is intended. Nevertheless, the following guidelines will produce a structure and order that is suitable for most purposes. It is suggested that the report, after the title page, should deal with the various sections in the order as set out below :

(i) Title page

The first page, just inside the report cover, usually repeats the title of the report, gives the list of authors and the name of the institution from which the survey was conducted as well as the period over which the study was carried out.

(ii) Acknowledgments

The next page is usually headed : **Acknowledgments,** and expresses thanks and appreciation to the various bodies and persons who have supported the survey. The importance of acknowledgments should not be underrated. Individuals, as well as organistaions, easily feel aggrieved (hurt) if their encouragement and support is not mentioned. It is better to acknowledge too many and too much than to omit or understate the contributions made by others. The page of acknowledgments, therefore, is best placed in a prominent position in the main report. The acknowledgments usually commence with sentences of appreciation for encouragement and support received from government officials or service departments. Any financial support given must be clearly stated although it is not usual to quote (mention) the actual sum of money received. The next acknowledgments, where they apply, are to other institutions such as universities, professional bodies or

business organisations that may also have provided advice and support. The final acknowledgments are usually to individuals who have helped with the study either by encouragement, expert advice or actual participation in the planning and execution of the survey.

(iii) Correspondence address

Readers with a particular interest in the subject of the study may wish to have further information and clarification. The wish is to be expected because the report, if it is not to become excessively long, must curtail (restrict) the detail and depth to which the survey methods, findings and conclusions are discussed. Hence enquiring readers must have a name and an address, usually the name of the survey organiser and the address of the institution at which he or she works, to which they may write for further information and clarification.

The correspondence name and address should be placed in a fairly prominent position in the report. Just below the Acknowledgments is often a suitable place.

(iv) List of contents

The list of contents, which best appears on a new page following the Acknowledgments, serves two main purposes :

(a) to inform the first time reader what subjects and topics are discussed in the report.

(b) to serve as a reference so that the reader can later find particular sections and tables quickly.

The contents list consists of a wide column in which the Section and Subsection titles are listed and against each of which their page number is given. The use of the contents list is made easier if the main sections are numbered and the subsections within each main section are again enumerated in a similar way. The titles used in the contents list, for ease of cross-reference and identification of topics, should correspond exactly with the section and subsection headings used in the report. The contents list given below is fairly typical :

CONTENTS

3. Writing the Report

There are many ways of writing and setting out a survey report, depending largely on the style and personal preferences of the author, and the readership he or she has in mind. Nevertheless, whatever method of presentation is adopted, it should aim to :

(i) be clear and readily understood

(ii) be pleasant to read and well laid out

(iii) arrange chapters and sections logically, with some means of cross referencing sections, tables and diagrams.

(iv) balance the amount of text and discussion on the one hand against the tables, diagrams and data on the other.

(v) stress appropriately the more important aspects and conclusions of the study.

(vi) avoid unnecessary detail, excessive length and repetition.

(vii) ensure that technical terms and unfamiliar expressions are explained in the text, in a footnote, or defined in a glossary. Non-essential technical terminology should be kept to a minimum or avoided altogether.

These guidelines are given further consideration under separate headings.

(i) Readability

The author's personal style is perhaps the most important single factor contributing to readability. Some writers have a natural way of writing that is both clear and pleasing. Not everyone is so fortunate. However, everyone can attend to the following points, which will improve readability :

(a) Avoid long sentences. Long sentences nearly always cover more than one idea, topic or condition. It is possible, in nearly all cases, to break up a long, involved sentence into several shorter ones that are clear and easily understood.

(b) Use simple words where possible, rather than unusual words. Avoid slang as well as phrases that are not in common usage.

(c) Divide the text into paragraphs, preferably short, where each paragraph concentrates on a single aspect or idea. The division into paragraphs helps the reader to follow what is being said because his mind concentrates on only a single theme whilst reading that paragraph.

(d) The last sentence of a paragraph can often be made to suggest, or lead into, the topic of the next paragraph thereby giving the text a feeling of logical continuity. The feeling of continuity is increased if the first sentence of the next paragraph takes up the topic alluded to by the previous sentence.

(e) Technical terms and expressions should be kept to the minimum possible. Where such terms are unavoidable, they should be explained to the reader in the text, or in a footnote, the first time the technical terms and expressions occur in the report. Alternatively, if the list of technical terms is fairly long, a small glossary (mini-dictionary) of the words can be included in the report, usually as an appendix.

(f) Repetition should be avoided, as it lengthens the report unnecessarily, adds no new information and bores the reader. Repetition applies not only to saying essentially the same thing again within the same paragraph, but applies equally to re-discussing a topic, or some aspect of it, that has previously been dealt with in another section. There may, of course, be some particular aspect that is so important that it will be referred to again elsewhere in the report. Such deliberate repetition, to emphasise a central theme, may well be justified. Nevertheless, even crucial points should not be repeated unnecessarily.

(ii) Logical arrangement

A clear, logical order in which the topics are discussed is immensely helpful to the reader, for two reasons. Firstly, a logical sequence is much easier to understand and to follow, and secondly, because the serious reader will need to refer back to earlier sections of the report as he or she studies it. Referral to other sections is greatly facilitated (made easier) if the report has a consistently logical order.

The following procedure will help achieve a logical order and systematic development of ideas within the report :

(a) Make a list of the principal sections (chapters) that are essential to the report. Only a brief descriptive title is needed for each of the sections at this stage.

(b) Next, order the main sections in the sequence in which they should appear in the report. Authors should then ask themselves whether there is any information that will be discussed later in the report, and which is needed by the reader before he can understand the first chapter. If not, then the chosen chapter is suitable for starting the report. If, however, later topics are required to understand the first section, then the chosen chapter may not be the best chapter with which to start; probably some other section should come before it. Alternatively, the contents of the chapter should be expanded so as to contain within it all the information required for its understanding.

If the above procedure is followed for each of the chapters in turn, the final arrangement will be systematically and logically ordered. The reader will now be able to read through the report, chapter by chapter, without the danger of becoming confused by topics for which he has not been prepared by earlier sections.

Most reports, if they are comprehensive, will subdivide the chapters into a number of sub-sections. To ensure that the subsections within a chapter also follow a logical order, the same procedure should be applied, i.e. for each chapter, make a list of descriptive titles for its sub-sections and then order them so that what the reader has already read, has always prepared him for the next sub-section.

(iii) Balanced presentation

Some information, e.g. discussions and conclusions, is best conveyed in words and text whilst other information is better understood when expressed as tables, statistical indices and diagrams. Readers will appreciate the report most when text, statistical data, tables and diagrams are placed in proper sequence and are kept in balance. The text can then refer to adjacent tables or diagrams which, in turn, are explained by the text, all of which is helpful to the reader. Proper balance

between text, tables and diagrams makes the report easy to read, as well as clear and informative. It also results in considerable economy of space, as text without supporting tables and diagrams tends to become long, verbose (wordy) and repetitive. However, it may not be practical, or desirable, to put all the tables and diagrams into the body of the report; * to do so may make for heavy and uninteresting reading. A well judged balance is required as to what goes into the main report and what is best put into the appendices. In a really well written report, the reader should be able to understand, on first reading, the main findings and conclusions, without having to refer to the appendices. A study of the details and additional information, which are usually put into the appendices, should only be necessary on second reading and only for those with particular interests.

Most readers, even if highly experienced and intelligent, cannot absorb too much detail on a first reading of a lengthy report. The writer has the responsibility to separate out the essential information and the most important conclusions. If the writer fails to do this, then the report is unbalanced and will lose the interest and concentration of all but the most determined and dedicated readers. Instead of becoming widely read, the survey report will, at best, remain with a small circle of specialists. Specialists are not always the people with the influence or the resources to help implement the recommendations of the report. A poorly written and unbalanced report has been the death knell of many a survey.

Finally, the temptation must be resisted to report every minor fact, occurrence and detail. Readers have neither the time, nor the interest, to wade through pages and pages of unimportant matters. Here is where the writer must exercise his judgement and severe restraint, expanding on the really important data and issues whilst condensing the less important; some matters do not require reporting at all. The appendices, too, must not be allowed to become bulky and overloaded with unnecessary and irrelevant information.

---

* The body of the report comprises all the main text, but usually excludes the initial introduction and the later appendices.

(iv) Cross-referencing

A method of cross-referencing is created whenever sections, subsections and possibly even the paragraphs, are identified (marked) by a consistent enumeration system, numerical or alphabetic. The List of Contents, on page 8, provides an example, the first part of which is reproduced below.

Section A : Coding **Page**

Proper cross-referencing is a great help to both the report writer and to the serious reader. It provides the writer with a flexible means of avoiding repetition; he needs only to quote the cross-reference code or page, to refer the reader to another part of the report. Likewise, the reader can, by looking at the List of Contents, see where in the report he can find the sections in which he is interested.

There are many systems of cross-reference. Most reports of moderate length require only a simple, consistent and logical system of identification. Whatever system is chosen, it must be used consistently throughout the report.

The three components of a cross-reference system are :

(a) A list of contents that lists the section and sub-section titles, the cross-reference enumeration code and the corresponding page number.

(b) The numbering of the report pages.

(c) The obligatory (essential) appearance in the text of all, or part of the appropriate enumeration code, which is usually in the margin next to, or as a prefix to, the corresponding title or sub-title.

The cross-referencing used in this booklet has four levels :

First Level : The Section code A, B, C, D or E and the Appendices.

Second Level : Within each Section, there are numbered sub-sections; for instance, Section A has five sub-sections. The section called "Appendices" has five, each being an appendix on some special topic and numbered appropriately.

Third Level : Some of the sub-sections have further sub-sections which are enumerated by lower case (small letter) Roman numerals such as (i), (ii), (iii).
For instance :

4. Extracting the Information :
   (i) Tally chart extraction
   (ii) Summary chart extraction.

Fourth Level : An example is given by the entry :

(1st Level) **Section A : Coding**

(2nd Level) 2. Coding Methods

(3rd Level) (i) Coding closed questions

(4th Level)
- (a) Closed questions whose options are mutually exclusive
- (b) Closed questions whose options are not mutually exclusive
- (c) Coding of priorities and pathways.

Cross-referencing allows the writer to refer, not only to a particular page, but also to whole sections and sub-sections very specifically. Thus a reference given as : (A, 2(i), (b)) refers to the whole sub-section : "Closed questions whose options are not mutually exclusive". Thus, for both the writer and the reader, the code (A, 2(i), (b)) is a convenient and efficient shorthand, making unnecessary the writing out of titles and sub-titles. The reader, given such a cross-reference, needs only to look up the List of Contents to find the title of the sub-section and the page on which it starts.

Although a proper cross-reference system is convenient, it should not be made unnecessarily complex. Anything beyond the four levels, as demonstrated above, is likely to be unrewarding.

(v) Appendices

The following are the principal reasons for having appendices :

1. To decrease the bulk and increase the readability of the body of the report.*

---

* In this section, the body of the report will, for simplicity, be referred to as the main report.

2. To give the writer flexibility to concentrate on the themes of greatest importance, whilst relegating themes of lesser interest to the appendices.

3. To make it easier to maintain a balanced main report. An excessive number of tables, diagrams and other supporting information may distract the reader; too much information may confuse rather than enlighten.

4. To provide space for topics that are of purely specialist interest and of only limited interest to the majority of readers.

5. To provide space for explanatory information that may be unnecessary for better informed readers.

Appendices augment the main report, as the above list shows. The main report is incomplete without the appendices and may, in parts, be difficult to follow in their absence. Appendices are, therefore, important and must be written with the same care and attention as the body of the report. They must, however, not be allowed to become too long, and technical terminology, if used, must be explained. They should be specific to a particular issue, theme or topic. If several topics and issues need discussion, it may be better to deal with each in a separate appendix. Each appendix should, as far as practical, be single minded and concentrate on a narrow, specific topic; an appendix should not be used for a wide ranging review or broad discussion.

The appendix is a marvellously flexible tool if used wisely.

Section E : Some Concluding Remarks

The analysis of survey data, the presentation of the results and the final report writing are substantial tasks that undoubtedly benefit from imagination, professional knowledge and experience. This should not, however, deter the first time survey organiser; after all, every specialist started with his or her first study. These rules are quite simple and require only that the organiser be willing to discuss his problems and plans with others throughout the study.

Rule 1. Discuss your survey plans with your medical and professional colleagues before you start and also during the unfolding of the study. Their comments and criticisms can be invaluable.

Rule 2. Where possible, get in touch with a statistician or survey specialist with whom you can discuss the sampling plan, the questionnaire design and analysis as well as other aspects of the study.

Rule 3. Persuade colleagues and senior collaborators in the study to read and comment on the early drafts of your report.

Rule 4. Check that the dissemination of the report is as complete as possible; ask your colleagues for ideas as to who else should be on your circulation list.

Seeking advice and comments, whilst nearly always extremely helpful, will often bring criticism and conflicting recommendations. Criticism is not always easy to accept, but do not let that deter you from giving it careful consideration; after all, the criticism may be valid. In surveys, several ways of dealing with a problem are often possible. Inevitably, those advising will not all give precisely the same advice; it may even be contradictory. In the end, the author of the study and of the report must make up his own mind and have the courage and determination to proceed as he thinks best. Apart from their scientific aspects, surveys are influenced by the views, imagination and personality of the originators. Surveys, like other creative activities, will always, to some extent, reflect the character and purpose of those who conduct them. That is part of their fun !

HEALTH CELL
TM-110
08038 N89

Appendix I

Description of Five Sampling Designs

(i) List sampling

List Sampling consists of attaching a number to every study unit in the survey population and then drawing a random sample of the numbers. Study units whose numbers coincide with the numbers drawn are then included in the survey.

(ii) Numbered tag sampling

Numbered Tag Sampling consists of issuing (giving) every person a numbered tag as he/she comes to a clinic or applies for some service. Only those persons whose tag number ends in previously agreed digits (numbers) are included in the study.

(iii) Stratified sampling

The whole survey population is divided into groups or strata in such a way that within each stratum the study units are more alike than they are in the survey population as a whole. Separate areas or institutions in which similar social or health conditions exist, can also be considered as strata which, when taken together, must cover the whole population. A separate sample is taken from each and every stratum using the above list sampling procedure.

(iv) Cluster sampling

Cluster Sampling consists of groups or clusters of sampling units enclosed in an easily recognisable boundary. When forming clusters, the study units within a cluster do not need to be similar. All clusters should contain, as far as is practical, approximately the same number of study units. A random sample of the clusters is chosen by list sampling and all the study units within the selected clusters are examined or interviewed.

(v) Two-stage sampling

Clusters are formed as for cluster sampling and a random sample of the clusters is chosen. A list is then made of all the study units within the selected (sample) clusters. By the list sampling method, a sample of study units is then drawn from each of the selected (sample) clusters.

Appendix 2

Estimating Totals for Stratified, Cluster and Two-Stage Sampling Designs

In order to summarise survey information and to make it more comprehensible, many different statistical quantities can be estimated using the sample results. Amongst the most commonly used are : percentages, averages, totals, the median, standard deviation and so on. These statistical quantities, often referred to as parameters, are in the main easy to calculate for simple random sampling designs such as list or numbered tag sampling. These same parameters are still of primary interest in more complex sampling designs, but their computation becomes more involved and difficult. The exact method of calculation is determined by the details and type of survey design used. A survey statistician should be consulted under such circumstances.

A knowledge, or at least a reliable estimate, of community totals, is perhaps the single most useful parameter for purposes of health services planning and resource allocation. Planning school, hospital or maternity services requires respectively an estimate of the total number of pupils, the number of acute admissions in a year or the annual number of pregnancies and births for which provision has to be made. The following will therefore describe in some detail how to estimate totals from three common survey designs :

1. Stratified sampling
2. Cluster sampling
3. Two-stage sampling.

Readers are urged to refresh their memories of the sampling procedures used with the above three designs.* The important concept of a study unit is defined as the smallest unit chosen by the sampling method and which the interviewers must ultimately visit or examine. In many surveys the study unit is the household.

---

* See Appendix 1 for the definitions and Booklet 2 in this series on Sampling.

Estimating Totals :

I. The Stratified Sampling Design

In the stratified sampling design, a simple random sample of study units is drawn from each of the strata into which the community or survey area has been divided. The exact number of study units in each stratum must be known beforehand, or must be counted before sampling begins. To make the computations easier to follow, we introduce the following symbols :

(i) K is the number of strata covering the community or survey area

(ii) N_1, N_2, N_3, ..., up to N_K are the **number of study units** in stratum 1, stratum 2, stratum 3, ..., up to stratum K respectively.

(iii) n_1, n_2, n_3, ..., up to n_K are the number of study units taken into the survey from stratum 1, stratum 2, ..., up to stratum K, respectively, i.e. n_1, n_2, ..., n_K are the stratum sample sizes.

(iv) t_1, t_2, t_3, ..., up to t_K are the total number of items, e.g. the number of persons over age 60 in a household, **as found in the samples** taken from stratum 1, stratum 2, ..., up to stratum K, respectively.

To estimate the required total, we proceed as follows :

Step I :

Compute, for **each** stratum, the **sample total.** It is easily done by counting how many of the items being investigated were present in the sample chosen from each stratum, thereby giving the values of t_1, t_2, ..., up to t_K.

Appendix 2

Step II :

Calculate for **each** stratum, the estimated **stratum total,** which is given by the expression :

$(\frac{t_1}{n_1})\ N_1$ for the first stratum,

$(\frac{t_2}{n_2})\ N_2$ for the second stratum,

and similarly for the other strata.

Step III :

The estimated **community total** is then found by adding together all the totals calculated under Step II.

Example :

A survey covered three villages, each of which was treated as a stratum, and which contained 102, 57 and 161 households, i.e. $N_1 = 102$, $N_2, = 57$ and $N_3 = 161$. Random samples of 11, 6 and 17 households were drawn from the villages, i.e. $n_1 = 11$, $n_2 = 6$ and $n_3 = 17$. As described in the text (see page 29), the survey questionnaires would be placed in separate piles according to the stratum from which they came. In this survey the total number of persons aged 60 and over found in each of the stratum questionnaire piles was 16, 9 and 24, which are the values for t_1, t_2 and t_3 respectively. The estimated sub-totals for each stratum (of over) are then given by :

$$(\frac{t_1}{n_1})\ N_1,\ (\frac{t_2}{n_2})\ N_2,\ \text{and}\ (\frac{t_3}{n_3})\ N_3$$

yielding :

$$(\frac{16}{11}) \times 102,\ (\frac{9}{6}) \times 57,\ \text{and}\ (\frac{24}{17}) \times 161$$

Rounding the calculations to the nearest integer (whole number), we obtain 148, 86 and 227 persons aged 60 and over in the three strata respectively.

Finally, the estimated total number of persons aged 60 and over for the whole community of three villages is the sum of the sub-totals for all the strata :

148 + 86 + 227 = 461 persons aged 60 and over.

II. The Cluster Sampling Design

In a cluster sampling design, the community or survey area is divided into M clusters which should, as far as possible, contain approximately the same number of study units. The study units within each cluster should preferably be a typical mixture of the various kinds or types of units commonly found in the community. A random sample of m clusters are chosen from the M clusters.

When the completed survey questionnaires are all collected, they are sorted into m piles, each pile corresponding to the questionnaire coming from a particular cluster. The total items e.g. number of children under the age of 3 years, found in each cluster is recorded and denoted by t_1, t_2, up to t_m for the m^{th} cluster in the sample. These m sub-totals are added to give a sample total which we denote by T.

The estimate for the total of the whole community is then given by the expression :

$$\text{Estimated Community Total} = \left(\frac{T}{m}\right) M.$$

Appendix 2

Example : In a fair sized town, the area under the administration of the town councillors is divided into 38 districts. In this example, each district is treated as a cluster, so that M, the total number of clusters in the population, is 38. A random sample of 10 of the districts was drawn, i.e. m = 10. Each household in the 10 sample districts was visited over a short period. One of the questions asked at each household, and recorded, was the number of children aged up to three years. After sorting the questionnaires into piles corresponding to each of the clusters, it was found that the number of children under age 3 for the 10 sample clusters was :

61 87 34 117 92

54 97 122 88 75

The total for all the ten sample clusters is therefore the sum of the ten sample values = 827 = T.

The estimate for the total number of children in the town under the age of three is then found by using the expression :

$$\left(\frac{T}{m}\right) M = \left(\frac{827}{10}\right) \times 38 = 3143, \text{ to the nearest integer.}$$

Note :

Although a cluster sampling design is often advantageous as regards the survey organisation and supervision, it is not always an efficient sampling procedure. Cluster sampling, more than other survey designs, can, on occasion, give unrepresentative results and yield misleading estimates of totals, averages, percentages and other statistical parameters. Where the survey organiser has the choice, he should consider a stratified or two-stage sampling design instead of cluster sampling.

III. The Two-Stage Sampling Design

In a two-stage sampling design, the community or survey area is first divided into M clusters or districts.*

It is an advantage if the clusters are of roughly the same size, i.e. they all contain approximately the same number of study units. At random, m of the clusters are chosen. The number of study units within **each** of the m sample clusters must be counted. Let the number of study units within the sample clusters be denoted by N_1, N_2, N_3, ..., up to N_m for the last of the m clusters drawn into the sample. Then from **each** of these clusters we randomly draw a sub-sample of study units and denote the number of study units drawn by n_1, n_2, n_3, ..., up to n_m, i.e. a sample of n_1 is drawn from the N_1 study units in the first cluster in our sample, and similarly for the other clusters.

To estimate the total for the whole community we proceed as follows :

Step I :

Determine from the completed questionnaires, after they have been put into m piles, corresponding to each of the m sample clusters, the total number of relevant items found in each of the m piles of questionnaires. Typical examples might be the total number of disabled persons or the number who have malaria, found in the sample clusters. Denote these sample totals by :

t_1, t_2, t_3, ..., up to t_m for the m^{th} sample cluster.

---

* The important difference between cluster sampling and two-stage sampling is that in cluster sampling **every** study unit is examined or visited within the chosen clusters. In the case of two-stage sampling, only a random sub-sample of units is drawn from the selected clusters.

Appendix 2

Step II :

Estimate the **sample cluster** totals by

$$T_1 = \left(\frac{t_1}{n_1}\right) N_1, \text{ for the first cluster,}$$

$$T_2 = \left(\frac{t_2}{n_2}\right) N_2, \text{ for the second cluster,}$$

and similarly for all the m sample clusters.

Step III :

Finally, sum all the T_1, T_2, T_3 up to T_m and denote the total by T. The figure for T gives the estimated total number of such items for the survey sample.

Step IV :

The total for the whole community is then estimated using the expression :

$$\text{Estimated Community Total} = \left(\frac{T}{m}\right) M.$$

Example :

A community, known to be exposed to malaria, consisted of 83 small villages spread out along a river area and its tributaries. A two-stage sampling scheme was devised to estimate within the community, the number of cases who had suffered a severe attack of malaria during the past 12 months. Here M, the number of clusters (villages in this case), is 83.

A random sample of 20 villages was chosen from the 83, i.e. m = 20, and a list made of the households in each of the 20 villages chosen to be part of the survey.

Appendix 2

Using the lists of households, a random sample of 1 in 5 of the households was chosen from each of the 20 survey villages. The sample size was rounded up where 1 in 5 did not give a whole number. For instance, in the example given below, the first village consisted of 21 households; a one in five sample would be 21/5 = 4.2. Because 4.2 is not a whole number, the sample size was **increased** to the next integer, i.e. increased to 5, as shown in column 3. Each of the sample households was then visited and the number of active malaria sufferers recorded. To ease the task of estimating the community total, the results were set out as shown in the table overleaf.

Hence the estimated total of active malaria sufferers for the whole community is :

$$\left(\frac{T}{m}\right) M = \left(\frac{420}{20}\right) \times 83 = 1743$$

taken to the nearest whole number.

The estimated total may, for planning purposes, be taken as 1750.

Appendix 2

Sample Village	No. of Households in village	No. of Households visited	No. of malaria cases per Household visited	Number of Malaria cases found	Estimated Cluster Totals (to nearest integer)
1	21	5	1,0,2,3,0	6	6/5 x 21 = 25
2	16	4	2,1,1,0	4	4/4 x 16 = 16
3	15	3	1,1,1	3	3/3 x 15 = 15
4	26	6	2,0,0,1,1, 1	5	5/6 x 26 = 22
5	47	10	3,0,1,0,2, 4,0,1,0,1	12	12/10 x 47 = 56
6	27	6	3,1,1,0,0, 0	5	23
7	32	7	1,0,0,0,1, 1,1	4	18
8	18	4	1,1,0,0	2	9
9	62	13	0,0,0,0,1, 2,1,0,0,1, 1,1,1	8	38
10	40	8	1,2,2,3,0, 1,1,0	10	50
11	13	3	1,1,1	3	13
12	10	2	0,0	0	0
13	22	5	0,0,0,1,2	3	13
14	31	7	2,0,0,1,1, 2,1	7	31
15	19	4	1,1,2,2	6	29
16	34	7	0,0,0,1,1, 2,0	4	19
17	15	3	0,0,1	1	5
18	20	4	1,0,0,1	2	10
19	35	7	0,0,0,0,0, 1,1	2	10
20	18	4	2,1,0,1	4	18
					T = 420

Appendix 3

Estimating the Median and the Quartiles from Frequency Tables

Estimating the median and the quartiles for **small** samples, say up to 50, is best done by arranging the data in ascending order and then proceeding as explained in the text, see pages 52 and 60. However, for large samples it is tedious to set out the data in ascending order and a quicker method is to estimate the median and the quartiles from the percentage frequency tables. The method will be illustrated using the 448 systolic blood pressures obtained during the Edinburgh-Fife Heart Study.

Step I :

Construct the frequency table in the usual way; the class intervals need not be of equal size. Next, calculate the percentage frequencies and then successively cumulate (add) the percentages as shown in the fourth column of the table below.

(i) Systolic B.P. (mm Hg)	(ii) Frequency	(iii) Percent Frequency	(iv) Cumulative Percentages
100 - 119*	62	13.8	13.8
120 - 139	201	44.9	58.7
140 - 159	127	28.3	87.0
160 - 179	40	8.9	95.9
180 - 199	14	3.1	99.1
200 - 219	3	0.7	99.7
220 - 239	1	0.2	99.9
Totals :	448	99.9 **	—

* Some tables indicate the class interval by the convention 100 < 120 instead of 100 – 119, etc. as done here. The two methods are equivalent.

** Note the small rounding error that gives a total of 99.9 instead of 100.0.

Appendix 3

Note :

The second percentage in column (iv) is obtained by adding 44.9 to 13.8 = 58.7; the third percentage is obtained by adding 28.3, i.e. 28.3 + 58.7 = 87.0, and similarly for the remaining cumulative percentages.

The cumulative percentages convey the following information :

(i) 13.8 percent of all the observations are less than 120 mm Hg, the start of the second interval.

(ii) 58.7 percent of all the observations are less than 140 mm Hg, the start of the third interval.

(iii) 87.0 percent of all the observations are less than 160 mm Hg. Similar interpretations apply to the remaining cumulative percentages.

Bear in mind the definitions of the quartiles and median, namely :

1. Q_1 is a value such that 25% of the observations are less than Q_1;
2. the Median (Me) is a value such that 50% of the observations are less than Me;
3. Q_3 is a value such that 75% of the observations are less than Q_3. In the above example, Q_1 and Me both lie within the interval 120 – 139, whilst Q_3 lies in the next class interval, 140 – 159.

Step II :

We now concentrate only on the class interval into which the corresponding Q_1, Me and Q_3 falls.

Next, note down the following values :

(i) the starting value of the class interval with which we are concerned. Call the starting value X.

(ii) the length of the class interval. Call the length L.

(iii) the difference between the quartile or median percentage and the cumulative percentage at the start of the interval. Call the difference between these two percentages D.

(iv) the percentage frequency falling within the interval. Call the percentage P.

Then Q_1, Me and Q_3 are each calculated by the formula (expression) :

$$X + (\frac{D}{P})\ L$$

Example :

Using the systolic BP example above, for Q_1, we have :

X = starting value of the class interval into which Q_1 falls
= 120

D = 25 – cumulative percentage at the start of the class interval; the 25 here corresponds to the 25% value used in the definition of Q_1.

= 25 – 13.8 = 11.2

P = percentage frequency in the class interval

= 44.9

Appendix 3

L = Length of class interval = 140 – 120

= 20 mm Hg.

Hence $Q_1 = 120 + (\frac{11.2}{44.9}) \times 20$

$= 124.99 = 125$ approximately.

To estimate Me, which in the above example happens to fall into the same class interval, D now becomes 50 – 13.8 = 36.2. The value of 50, used in calculating D, corresponds to the 50% value used in the definition of the Median. All the other values remain as above because Me, in this example, happens to be in the same class interval as Q_1.

Hence :

$$\text{Median} = 120 + (\frac{36.2}{44.9}) \times 20 = 136 \text{ approximately.}$$

Finally, for Q_3, which corresponds to 75% of observations and which, in the example, falls into the next class interval, we have :

X = 140 (starting value of the interval into which Q_3 falls)

D = 75 – 58.7, where 58.7 is the cumulative percentage at the start of the class interval.

= 16.3

P = 28.3, the percentage frequency for the interval

L = 160 – 140 = 20

$$\text{Hence } Q_3 = 140 + (\frac{16.3}{28.3}) \times 20 = 152 \text{ approximately.}$$

In the example almost 45% (44.9) of all the observations fall into the second interval and almost 30% (28.3) into the third class interval. The calculations of the quartiles and medians would be more precise if, in such instances, smaller class intervals were used for that part of the distribution.

In the example, it would have been better to replace the two intervals 120-139 and 140-159 by four smaller intervals 120-129, 130-139, 140-149 and 150-159 and to use them to construct the frequency table. An interval length of 20 mm Hg seems suitable for the class intervals before 120 and after 160 mmHg. It has to be recognised that the first time a frequency table is drawn up, using what seem to be sensible class intervals, we may obtain an excessive concentration of cases within one or two intervals. When that happens, these particular intervals can be divided into shorter intervals and the frequency table re-done, using the new, shorter intervals.

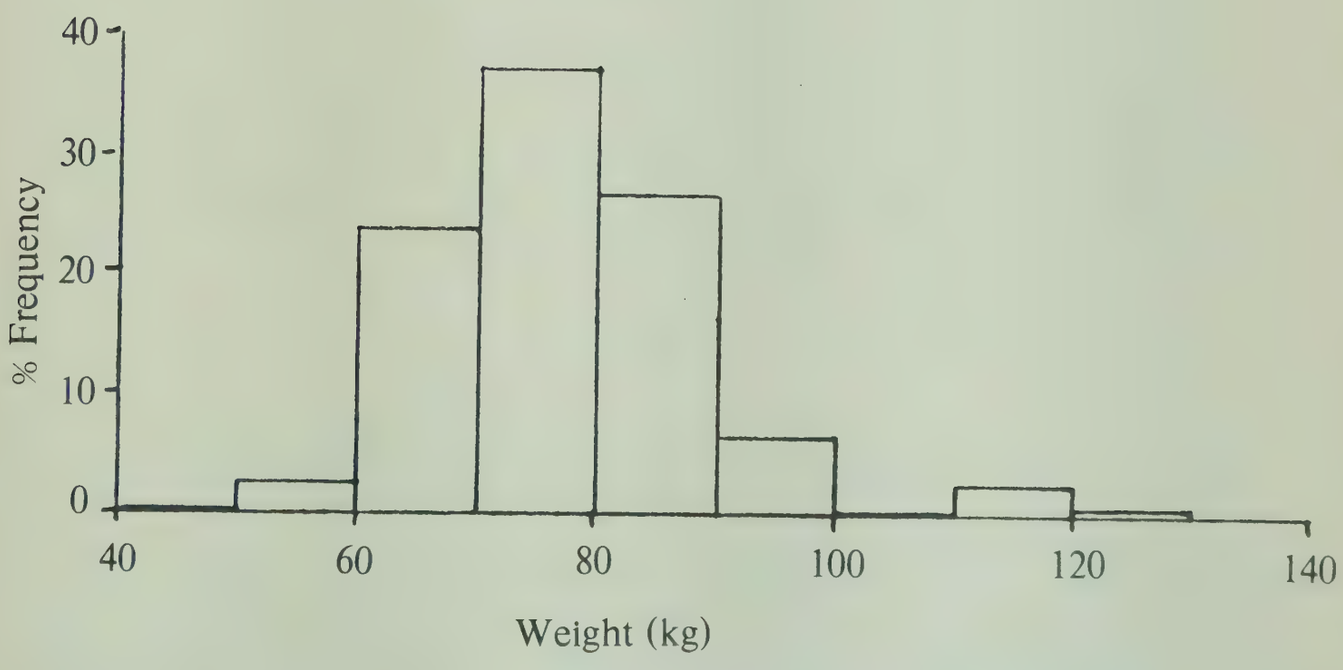
Appendix 4

When is the Median preferred to the Average ?

The average (mean) is the most frequently used measure of location. The average is in many ways the ideal typical value that lies somewhere near the centre of the distribution of values. The average is also favoured on purely theoretical grounds as it has admirable statistical properties.

However, the average is the ideal measure of location only when the histogram of the data is symmetrical or nearly so. The average would be the best typical value when the histogram has most of its observations, or its largest percentage frequency, near the centre of the distribution, as is the case in the histogram of adult male weights shown below :

Weight of 448 Edinburgh-Fife Males (Age 45-54) 1980



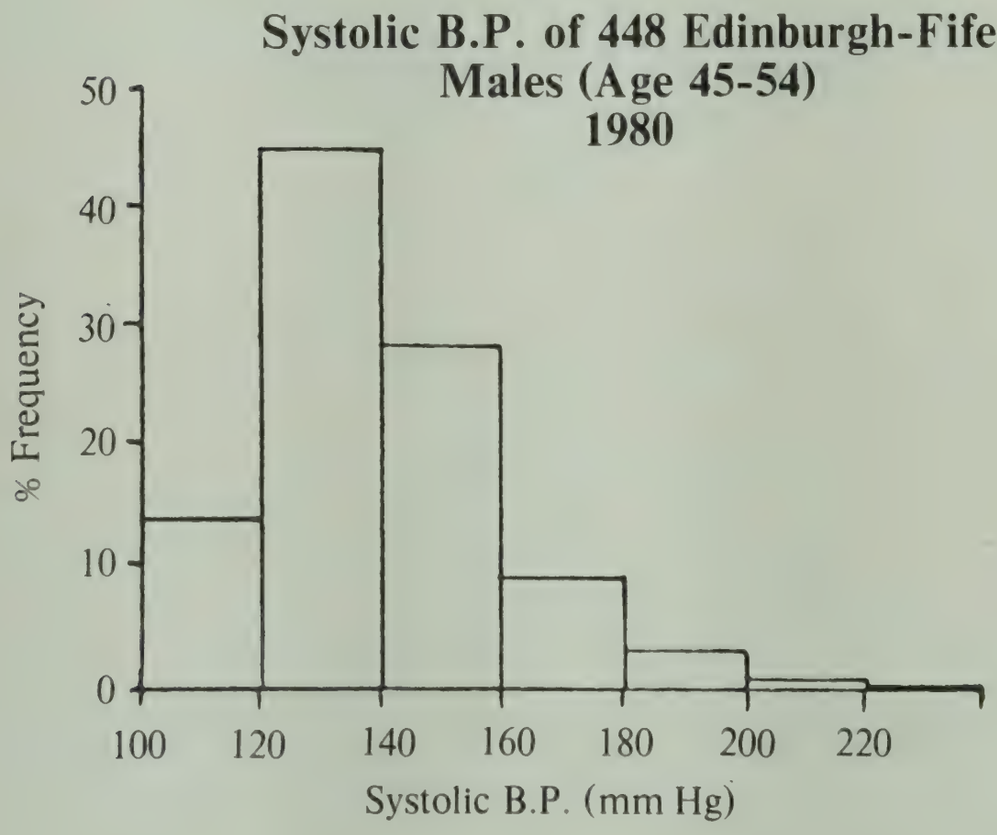
Appendix 4

Despite the general preference for the average as the ideal representative value, there are two situations where the median may be preferred :

(i) when the histogram is very skewed (asymmetrical), which is sometimes the case with certain physiological, economic and sociological variables.

(ii) when the shape of the histogram is not known, but it is suspected that the data has a very skew distribution. This applies particularly to small samples that are not large enough to draw the histogram, which would show how asymmetrical the distribution was.

An example of a skew histogram is shown below :



However, where the sample size is large, say 25 or more, the average will be preferred to the median unless the distribution is extraordinarily skew.

Appendix 5

The Variance, Standard Deviation, Standard Error and Confidence Intervals

There are serveral statistical estimates that are frequently used in the analysis of numerical measurements such as height, cholesterol level and blood pressure. The following three statistical estimates are particularly important :

(i) the variance and its square root, called the standard deviation

(ii) the standard error of a mean and of a proportion (or percent)

(iii) confidence intervals for a mean and for a proportion (or percent).

These statistical estimates are also important for the analysis of survey data, provided the data are in the form of actual measurements. The estimates apply to all types of survey sampling schemes, but their method of calculation can become complex. The calculations are straightforward only for list and numbered tag sampling, often called simple random sampling methods, and therefore they are the only calculations that will be explained here. For other sampling designs, the reader is advised to consult a survey statistician for the computation of the estimates and their application.

I. The Variance and the Standard Deviation

There are two methods for estimating the variance. The first method is used for small samples, say less than 50 observations. The second method is based on the frequency table, and is recommended for larger samples, say 50 or more.

1. Estimating the variance for small samples

To illustrate the method, we consider the results of a survey of community drinking water supplies where the measurement is the depth, in metres, down to the water level of ten wells :

1.4	0.8	3.4	2.7	1.9
3.9	4.2	1.8	2.4	2.1

The calculation of the variance proceeds as follows :

Step I :

(a) calculate the sum of all the values, i.e. add up all the values and call the sum A, giving for the above data :

A = 24.6

(b) square each of the sample values and add them up; let the result, called the sum of squares, be denoted by SS. Note that it is often easier to write down the value of the individual squares before adding them up. e.g. 1.4 x 1.4 = 1.96, for the first result.

We have, for the above data :

1.96	0.64	11.56	7.29	3.61
15.21	17.64	3.24	5.76	4.41

The sum of squares, SS = 71.32.

Step II :

Let the sample size be denoted by n; for the above data, n = 10.

The sample variance, denoted by V, is then calculated using the formula :

$$\text{Variance} = V = \frac{1}{n-1}\left[SS - \frac{AxA}{n}\right]$$

giving for the above data on well depths :

$$V = \frac{1}{9}\left[71.32 - \frac{24.6 \times 24.6}{10}\right] = 1.2 \text{ approximately.}$$

2. Estimating the variance for large samples

To illustrate the method, we consider the previously given frequency table of systolic B.P. found in a sample of 448 men aged 45-54 years in the Edinburgh-Fife Heart study (see columns (i) and (ii) below).

(i) Systolic B.P. (mm Hg)	(ii) Frequency f	(iii) Mid Point m	(iv) mxf	(v) mxmxf
100 - 119	62	110	6820	750200
120 - 139	201	130	26130	3396900
140 - 159	127	150	19050	2857500
160 - 179	40	170	6800	1156000
180 - 199	14	190	2660	505400
200 - 219	3	210	630	132300
220 - 239	1	230	230	52900
Totals :	448	–	62320 = A	8851200 = SS

Step I :

(a) set out the data as a frequency table as is shown in columns (i) and (ii) above. The class intervals **do not** have to be of equal length; the method is valid (applicable) for tables with unequal as well as equal class intervals.

(b) calculate the mid point value for each class interval; call the mid point value m and set it out as a third column (see column (iii) above).

(c) calculate the product m x f for each interval, i.e. multiply together, for each interval, the value of the mid point times the corresponding frequency for that interval; write the values down as a fourth column (see column (iv) above).

(d) mulitply each value of column (iv) by its corresponding mid point value, m, and set the results out as column (v), e.g. the first entry in column (v) for the systolic B.P. data is :

$$110 \times 6820 = 750200$$

Similarly for the other entries in column (v).

Step II :

(a) sum the values in columns (iv) and (v) to obtain the values for A and SS respectively (note the similarity to the method used for small sample sizes).

$$\text{The sample mean} = \frac{A}{n} = \frac{62320}{448} = 139.107 = 139 \text{ approx.}$$

(b) calculate the sample variance by the formula given previously :

$$\text{Variance} = V = \frac{1}{n-1} \left[SS - \frac{A \times A}{n} \right]$$

The systolic B.P. example gives :

$$V = \frac{1}{447} \left[8851200 - \frac{62320 \times 62320}{448} \right] = 407.25$$

$= 407$ approximately.

Appendix 5

Note :

(i) Although the second method may seem long, for large samples it is far shorter and quicker than the method given for small samples.

(ii) There exist several short cuts to the method, but the procedure given is the easiest one to remember and to use.

3. The Standard Deviation

The Standard Deviation (S.D.) is always given by the (positive) square root of the variance, by whatever method the variance is calculated.

Thus, for the depth of wells, we have :

S.D. = $\sqrt{1.2}$ = 1.1 (metres) approximately.

The standard deviation of the systolic B.P. is given by :

S.D. = $\sqrt{407.25}$ = 20.2 (mm Hg) approximately.

4. Application of the Standard Deviation

A common application of the standard deviation is the determination of an interval into which a certain percentage of **individual,** i.e. **single,** observations fall. Two intervals are most usually calculated, the 95% and 99% interval. They are calculated as follows :

(i) **the 95% interval :**
average – 2 x S.D. for the lower end of the interval, and average + 2 x S.D. for the upper end of the interval. This formula is usually written as :

Average ± 2.S.D.*

---

* Some textbooks use the more precise value of 1.96 instead of 2. The formula then becomes average ± 1.96 x standard deviation. In most applications it is sufficient to use the value of 2.

(ii) **the 99% interval :**
the procedure is exactly the same as for (i) above except that the expression used is :

$$\text{average} \pm 2.6 \text{ S.D.}$$

The 95% and 99% intervals are often used for trend curves. On pages 62 and 82 examples were given using the median and the quartiles Q_1 and Q_3 to establish intervals within which 50% of the observations would be expected to fall. Instead of the median and quartiles it is more common to plot the trend curve using either of the above expressions to show, on the graph, the intervals within which 95% or 99% of the observations are expected to lie.

Strictly speaking, the standard deviation should only be used in this way if the data have an approximately symmetrical bell-shaped distribution, i.e. have a symmetrical histogram, as, for example, shown on page 71 and in the first example in Appendix 4 (page 117).

II. The Standard Error

Everyone who does a survey, or who takes a sample, must decide what the sample results reveal about the population as a whole. If, during a survey, we determine the average daily calorie intake of a section of the adult population, then we know that the average so obtained is unlikely to be the exact average value for the whole of the population. Likewise, if during a survey amongst young mothers, we determine the percentage who weaned their first born at three months or less, then that percentage, as found in the study, is not likely to be the exact percentage for all young mothers. Statistics aims to give quantitative (numerical) answers to what the unknown population values are and to derive estimates from the sample results.

Appendix 5

The standard error is amongst the most important of the methods available to statisticians for making such generalisations. The standard error is a measure of the extent to which **sample estimates** will vary from one experiment to the next or from one survey to the next. Thus the **standard error of a mean** indicates the extent to which the mean will vary if a similar study were repeated. Similarly, the **standard error of a proportion** (or of a percentage) measures the variability of the sample proportion (or percentage) if the study were to be done again.

The expression (formula) for calculating the standard error will be different for each of the indices (estimates) we compute from the sample. Three important standard errors for simple sampling methods are given by the expressions :

(i) For an **average :**

$$\text{Standard error of mean} = \frac{\text{S.D.}}{\sqrt{n}}$$

where S.D. is the standard deviation and n is the sample size.

(ii) For a **proportion :**

$$\text{Standard error of proportion} = \sqrt{\frac{p(1-p)}{n}}$$

where p is the sample proportion and n is the sample size.

(iii) For a **percentage :**

$$\text{Standard error of percentage} = \sqrt{\frac{P(100-P)}{n}}$$

where P is the percentage and n is the sample size.

III. Confidence Intervals

The sample average, proportion and percentage are all estimates of what the value of the corresponding average, proportion and percentage for the population might be. The confidence interval is an interval within which we believe the population value will lie with a certain assurance of it being so.

The confidence interval, for **adequate sample size,** is always given by the expression :

Sample estimate $\pm$ K x (standard error of the estimate), where K = 2, approximately, for a 95% assurance of it being true,
or K = 2.6, approximately, for a 99% assurance of it being true.

Unfortunately, it is not easy to define what constitutes an adequate sample size; it very much depends upon the statistical estimate we are making. The following can, however, be taken as reasonable guidelines :

(i) when estimating averages, the sample size should be 20 or greater.

(ii) when estimating proportions between 0.1 and 0.9 and percentages between 10% and 90%, the sample size should be at least 100.

(iii) when estimating proportions or percentages below 0.1 or 10% respectively, samples larger than 100 are recommended; the smaller the proportion or the percentage, the larger the sample needs to be. The advice of a statistician should be sought for very small values. The same comments apply to proportions or percentages that exceed 0.9 or 90% respectively.

Appendix 5

Examples:

(a) From the above example of systolic B.P., obtained from a sample of 448 men aged 45–54 years (p. 121), the sample mean was 139.1 and its standard deviation equalled 20.2, both values being approximated to one decimal place.

The standard error of the mean is :

$$\text{Standard Error of Mean} = \frac{\text{S.D.}}{\sqrt{n}} = \frac{20.2}{\sqrt{448}} = 0.95 \text{ approximately.}$$

The 95% confidence interval for the unknown population mean, of which the survey of 448 men provides an estimate, is :

Mean ± 2 x Standard Error

which on inserting the previously calculated values, gives :

$$139.1 \pm 2 \times 0.95 = 139.1 \pm 1.9$$

i.e. we have a 95% assurance that the mean systolic B.P. for those men lies **somewhere** within the interval 139.1 ± 1.9, i.e. somewhere between 137.2 and 141.0 mm Hg.

(b) From the frequency table of 448 systolic B.P. we see that 58 men aged 45-54 years had a systolic B.P. of 160 mm Hg or higher,

i.e. $\frac{58}{448} \times 100\% = 12.9\%$ of the sample had a systolic B.P. of 160 or more.

The standard error for this percentage P is given by :

$$\sqrt{\frac{P(100-P)}{n}} = \sqrt{\frac{12.9(100-12.9)}{448}} = \sqrt{\frac{12.9(87.1)}{448}}$$

$$= 1.58$$

We then have a 99% assurance that for this particular population, the percentage with a raised systolic B.P. of 160 or more lies **somewhere** within the interval P ± 2.6 x standard error of the percentage. On inserting the above estimates for P and its standard error, we obtain :

$$12.9 \pm 2.6 \times 1.58 = 12.9 \pm 4.1$$

We therefore conclude that the population percentage lies somewhere within the interval : 8.8% to 17.0% with an assurance of 99%.

The results of health and community surveys become useful and meaningful only after the survey questionnaires and forms have been properly coded, the information extracted and analysed and the results reported. Straightforward statistical and graphical procedures are explained and demonstrated step by step, thereby enabling health care personnel to code, extract, analyse and interpret survey information. The importance of a clear, logical and interesting report is stressed and simple guidelines are given on how this may be achieved.

IMPRIMERIE DU GACHOT

75, ANCIENNE ROUTE
1218 GRAND-SACONNEX

TEL 022/98 07 51